ENGAGING WITH MUSIC RETRIEVAL

DANIEL BOLAND

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

APRIL 2015

Abstract

Music collections available to listeners have grown at a dramatic pace, now spanning tens of millions of tracks. Interacting with a music retrieval system can thus be overwhelming, with users offered 'too-much-choice'. The level of engagement required for such retrieval interactions can be inappropriate, such as in mobile or multitasking contexts. Music recommender systems are widely employed to address this issue, however tend toward the opposite extreme of disempowering users and suffer from issues of subjectivity and confounds, such as the equalisation of tracks. This challenge and the styles of retrieval interaction involved are characterised in terms of user engagement in music retrieval, and the relationships between existing conceptualisations of user engagement is explored. Using listening histories and work from music psychology, a set of engagement-stratified profiles of listening behaviour are developed. A dataset comprising the playlists of thousands of users is used to contribute a user-centric approach to feature selection. The challenge of designing music retrieval for different levels of user engagement is first explored with a proof of concept, low engagement music retrieval system enabling users to casually retrieve music by tapping its rhythm as a query. The design methodology is then generalised with an engagement-dependent system, allowing users to denote their level of engagement and thus the specificity of their music queries. The engagement-dependent retrieval interaction is then explored as a component in a commercial music system. This thesis contributes the engagement-stratified profiles and metrics of listening behaviour, a corresponding design methodology for interaction, and presents a set of research and commercial applications for music retrieval.

Acknowledgements

I would like to thank -

my wife, Estelle, for her patience,
my supervisor, Roderick, for his advice,
my colleagues, Mark and Ross, for their collaboration,
my friend, Melissa, for her kindness,
and you, reader, for your interest.

This work was supported by the University of Glasgow,
Bang & Olufsen and the Danish Council for Strategic Research.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Daniel Boland

Collaborations

In chapter 7 and chapter 8, Ross McLachlan and I jointly developed the initial concept for the prototype that was adopted and further developed to become the BeoSound Moment product. I contributed an engagement-dependent inferential model of music queries, to underpin the music recommendation. Ross incorporated design insight on the use of pressure, jointly developing the interface to include the pressure modality. A small design session was held by Ross to capture qualitative feedback. Subsequent analyses were performed independently.

In chapter 9, I contributed an engagement-dependent approach to blending Virtual Reality and some computer vision filters for use with a video pass-through head mounted display. Mark McGill implemented the study conditions in the Virtual Reality environment and conducted the study. Subsequent analyses were performed independently.

In appendix A, I implemented a pressure sensor and low-pass filter, the dynamics of which were optimised to maximise throughput in a Fitts' task. The experimental protocol was jointly designed with Ross McLachlan, who used his existing software to conduct the study. Subsequent analyses were performed independently.

Commercial Sensitivity

Some of the work described in this thesis occurred as part of collaborations with industry partners. Details of a commercially sensitive nature have been omitted, i.e., the details of the feature extraction for the playlist analysis in section 5.2 and the commercial context of the survival analysis in section 8.4.

List of Contributing Publications

Boland, D., McLachlan, R., and Murray-Smith, R. Engaging with mobile music retrieval. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI*, pp. 484–493. ACM (2015).

Boland, D. and Murray-Smith, R. Adapting recommendation to user engagement. In *Workshop on Principles, Techniques and Perspectives on Optimization and HCI, Conference on Human Factors in Computing Systems, CHI*. ACM (2015).

McGill, M., Boland, D., Murray-Smith, R., and Brewster, S. A. A dose of reality: Overcoming usability challenges in VR head-mounted displays. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI*, pp. 2143–2152. ACM (2015).

Boland, D. and Murray-Smith, R. Information-theoretic measures of music listening behaviour. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pp. 561–566 (2014).

McLachlan, R., Boland, D., and Brewster, S. Transient and Transitional States: Pressure as an Auxiliary Input Modality for Bimanual Interaction. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI*, pp. 401–410. ACM (2014).

Boland, D., McLachlan, R., and Murray-Smith, R. Inferring music selections for casual music interaction. In *Proceedings of the European Symposium on Human-Computer Interaction and Information Retrieval, EuroHCIR*, pp. 15–18. CEUR (2013).

Boland, D. and Murray-Smith, R. Finding *my* beat: personalised rhythmic filtering for mobile music interaction. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI*, pp. 21–30. ACM (2013).

TABLE OF CONTENTS

List of Tables and figures

1. INTRODUCTION

RETRIEVING music from a music collection is no longer as simple as picking a CD from a shelf. With music collections now mostly digital, they have grown to fill the increasing capacities of hard drives and music players. Where collections were once hand-curated by their owners, now control over the organisation of music is increasingly delegated to automatic systems. It would be impractical for listeners to manually scan a modern digital music collection for each retrieval, and so library management interfaces such as iTunes are used. This trend has continued with the launch of music streaming services such as Spotify offering access to millions of tracks, most of which the user is unfamiliar with and has little ability to organize. Users have limited means of exploring unknown content, with some 20% of tracks (4 million) on Spotify never having been listened to by anyone even once.[1] The shift from small user-managed collections to large digital libraries has introduced a need for novel music retrieval techniques and approaches to designing interactions with music. At present users are often faced with two extremes: the onerous task of manually selecting content (playlisting etc.) or yielding control to shuffle or radio playback.

[1] http://news.spotify.com/us/2013/10/07/the-spotify-story-so-far/ (16/12/14)

1.1 RECOMMENDER SYSTEMS

The issue of overwhelming choice in music was addressed by Celma (2010), his book on the *Long Tail* recommends eliminating some of the choices facing users through the use of personalised filters and recommender systems. Music recommendation enables users to easily access relevant content from a large music collection. Given that these collections often scale into the millions of songs, approaches to music recommendation typically rely on machine listening and collaborative filtering approaches rather than manual annotation. The performance of approaches taken in music recommendation and the challenges faced are explored in chapter 2, in particular the subjective nature of musical concepts such as genre, mood and similarity. There are a range of specificities at which music recommendations can be made, for example mood and genre recommendation allows users to retrieve music of a broad style, whereas similarity based recommendation aims to offer users music related to a seed track or artist, according to some distance function. Commercial music streaming services make extensive use of recommender systems to make their collections available to users, often presented using the metaphor of radio, automatically curated. Many music services make use of the music-listening technology built into the Echonest,[2] acquired by Spotify. Pandora is noted for its use of manual annotation (their 'Music Genome Project'),[3] collaborative filtering using data acquired from their relatively long time in the market, and giving users the option to provide relevance feedback in the form of 'thumbs up' and 'thumbs down'.

MORAL HAZARD

The issues of subjectivity and bias in music recommendation, detailed later in chapter 2, take on a significant prominence given the commercial and cultural importance of music consumption. Recommender systems are largely trained on music produced by major music labels, and may be confounded with the specific characteristics or production processes of this music. Music that deviates from the norms of a genre, or that is produced outwith a commercial music studio, may be incorrectly classified and thus unavailable to users via recommender systems. With users' retrieval of music increasingly reliant upon retrieval systems, there is the risk of niche or independent music being excluded from the market. It is all the more important then, to consider the implications of these recommender and retrieval systems becoming the gateway by which most listeners access music. Recommender systems allow users to delegate control to minimise the burden of choice. However, there is a need for this control to be negotiable, with users retaining the ability to determine how they access their music.

[2] `http://the.echonest.com/company/` (16/12/14)
[3] `http://www.pandora.com/about/mgp` (16/12/14)

1.2 ENGAGEMENT

Music retrieval varies from casual browsing and exploring to specific known-item retrieval. The user's engagement and investment in the retrieval process varies accordingly. The casual user does not need to dedicate as much attention to the interaction, and does not need to provide much input to satisfice. More engaged users are willing to invest time and effort into the retrieval, and being likely to have a more specific information need, they will provide the required amount of input. A review of existing work on user engagement and information need is given in chapter 3, developing the definition of engagement used in this work. The music retrieval behaviours described are linked to user engagement in chapter 4.

MOBILE USAGE

The amount of attention that users can invest in their music retrieval is dependent upon their context. Oulasvirta et al. (2005) showed that the cognitive resources available for interacting with a device are limited in mobile settings, with some reserved for attending to the user's environment. A user listening to music on a car radio, or on an iPod while jogging, will likely be less willing to carefully navigate a hierarchy of menus or type in a textual query. Mobile music retrieval interfaces thus often allow for casual input, e.g. shuffle-based playback offering users a random and serendipitous style of music retrieval (Leong et al., 2005). These limited, casual forms of music retrieval are often paired with more demanding, high engagement options, e.g. the menu hierarchy and textual retrieval seen in the Apple iPod[4] and Spotify.[1] Users are then forced to choose between these extremes of engagement. The mobile user, often unable to dedicate the attentional resources required to fully engage in their music retrieval, is then left disempowered in their music listening, with only the choice of random shuffle or yielding control to broad recommendation.

CASUAL SEARCH

When users interact with a music retrieval system, they may not be performing the type of well-defined retrieval task that Information Retrieval (IR) generally considers. The field has been able to rigorously characterise the performance of retrieval methods by defining tasks for evaluation tracks at the Textual Retrieval Conference (TREC). Music retrieval is instead often what Wilson and Elsweiler (2010) describe as *casual-leisure searching* – where a user's goal is not to satisfy an information need, but instead to fulfil a hedonic goal, such as when browsing or using serendipitous retrieval such as shuffle. Without a defined information need, concepts such as relevance become nebulous, complicating the task of evaluation.

[4] `http://www.apple.com/uk/ipodclassic/` (11/08/14)

1.3 DESIGN PHILOSOPHY

The latter chapters of this thesis involve the development of prototype systems to explore the design of music retrieval interaction in the context of user engagement. Building such systems implicitly involves making design choices as part of the interaction design, with the resulting systems and interactions conditioned upon these choices. It is therefore worthwhile to be explicit about the design philosophy applied. In his chapter on user-centred design (Norman, 1988, p. 187) emphasises the importance of the user's conceptual model and that interactive devices behave in a manner consistent with this model. He explains that designers must craft a system to inspire an appropriate conceptual model in the user, consistent with the system. This thesis 'closes the loop' in this communication of a model image, in that the systems developed are based on generative models of user behaviour. As users begin to use the systems, which are based on a priori assumptions of the user, the user will form their own conceptual model, which the system can then adapt its internal user model to.

SIMPLICITY & CONTROL IN INTERACTION DESIGN

It is desirable for interactions to be *simple* – operable by users without their needing to apply complicated conceptual models. Norman (1988) motivates this need for simplicity by pointing to the constraints on users' attention and memory. He argues that new technology should make tasks simpler, and that conceptual models for operating it should be learnable by users. While automation (such as music recommendation) is desirable for the resulting simplicity of the interaction, Norman warns of the danger of taking *control* away from users – users want to take varying levels of control at different times, and he even uses the example of music, contrasting the ease of radio listening against the control of playing the piano. Music is a compelling case for considering simplicity and control, with the iPod shuffle being the first example used by Obendorf (2009) in their book on minimalism in interaction design.

MANAGING MULTIPLE MODES OF ENGAGEMENT

This thesis considers the different levels of engagement a user may have when retrieving their music, and music retrieval systems have been designed to cater to specific levels in this range of engagement. It could be argued that a user's music retrieval needs are met by a combination of systems such as mobile shuffle on an iPod, browsing music websites online, and highly engaged textual retrieval in Spotify. The design philosophy outlined here motivates having a simple, consistent mental model that affords the range of interactions required by the user. This motivation is reflected in the design choices made in the development work in parts III and IV – the interactions are designed to employ as simple a mental model as possible, while still affording a range of music retrieval behaviours.

1.4 AIMS & OUTLINE

This thesis aims to address the issue of too–much–choice in music retrieval. A key theme in this work is considering the appropriate balance between recommendation and user control, with a user's engagement in the retrieval process as the determining contextual variable.

Part I - Background A broad outline of the field of Music Information Retrieval is given in chapter 2, highlighting how previous efforts have been limited by the subjectivity of music and the need for adaptive approaches. Given the broad usage and interpretation of the term engagement, chapter 3 provides a thorough review, identifies common themes and elucidates a robust definition applicable to Music Information Retrieval.

Part II - Understanding Music Listening Chapter 4 develops measures of music listening behaviour and engagement in music retrieval. The relationship between these measures and those from the music engagement questionnaires is then explored. Building upon the information-theoretic approach to defining engagement, chapter 5 applies entropy over music features to provide timelines of music listening style. Information-theoretic feature selection is used to identify music features that are relevant in capturing the organisation of playlists created by users.

Part III - Designing for Engagement in Music Retrieval Two systems are developed to explore how to design for user engagement in music retrieval. A low engagement music retrieval style is explored, influenced by the musical universals identified in chapter 2. In chapter 6, the example music retrieval system allows users to query music by tapping a song's rhythm. The issue of subjectivity remains, however, and the work shows how retrieval performance improves by adapting to users. In chapter 7, an engagement-dependent approach to music retrieval is introduced, adapting a recommender system's autonomy to the user's level of engagement. A generative model of user queries conditioned on engagement enables the inference of relevant music.

Part IV - Outlook & Beyond Music The concept of engagement-dependent music retrieval is adopted in the development of a commercial music retrieval system, which is described in chapter 8. The measures of music listening behaviour detailed in this thesis are then used in an exploratory evaluation of a prototype of the product. Finally, a consideration is made of how to generalise the approaches developed in this thesis. User interactions with music and other media using Virtual Reality Head Mounted Displays (VR-HMDs) share the issue of how to support users in attending to a retrieval interaction only insofar as desired. The engagement-dependent approach is applied to VR interaction in chapter 9 and evaluated in a typing study, showing how users can issue textual queries while remaining engaged with a virtual media experience.

1.5 REPRODUCIBLE THESIS

This thesis is written as an example of reproducible research, including literate programming – a single set of source files contains the code for generating all of the analyses, figures and text of this document. These source files are made available to allow others to reproduce the results of this work and to conduct further work. These files contain a mix of primarily LaTeX and R, with some occasional Python snippets. The code can be 'woven' into this thesis using the KnitR package. Information on the schema for the dataset as well as instructions for building this thesis and code from source are given in Appendix C.

DATASET

The analyses in this thesis make use of data collected in studies from participants as well as music metadata crawled from online services, which are made available in an SQLite dataset. The code used to scrape this dataset is available under the MIT open source license, and can be accessed at:
`http://www.github.com/dcboland/`

A large proportion of the data in this dataset has already been made available as the *Streamable Playlists with User Data (SPUD)* dataset, which includes playlists created by users, song metadata and some user listening histories. An interactive figure along with corresponding R code for importing data from *SPUD* and producing the analyses and plots is available at:
`http://www.dannyboland.com/spud/`

The information-theoretic feature selection in chapter 5 and the example system in chapter 7 make use of proprietary music feature data, and as such this data is not made available. The MoodAgent features are commercially sensitive, and thus not included in the *SPUD* dataset. Even with user data and the required computational power, large-scale music analyses require licensing arrangements with content providers, which often vary from country to country, presenting a serious challenge to academic MIR research. The adoption of commercially provided features has allowed the demonstration of the information-theoretic approach, and while the audio stream links are distributed, it is unlikely that many MIR researchers will have the resources to replicate all of these large scale analyses. The CoSound[5] project is an example of industry collaborating with academic research and state bodies to navigate the complex issues of music licensing and large-scale analysis.

[5] `http://www.cosound.dk/` Last accessed: 30/04/14

1.6 TECHNOLOGICAL CONTEXT

The work in this thesis took place in the context of the technologies, devices and services available. As with any work related to HCI, an effort must be made to adapt to the changing consumer reality. The timeline in Figure 1.1 outlines the technological changes that occurred around this work, which are reflected in the devices used for prototyping and evaluation.

Apr, 2010	**Apple iPad** - a milestone in the development of tablet devices, leading to their widespread adoption.
Sep, 2011	**Nokia N9** - mobile phone used for development and evaluation of the Query by Tapping approaches discussed in chapter 6.
Nov, 2011	**iCloud & Music Lockers** - Apple, Google and others launch cloud music services, allowing users to store and listen to large volumes of music online.
Feb, 2012	**Kinect for Windows** - motion sensing device, capable of skeletal tracking, capturing 3D scene geometry and recognising user gestures, used for the augmented virtual reality work in chapter 9.
Feb, 2013	**Microsoft Surface** - tablet device functionality matches laptops, used for development of the engagement-dependent system in chapter 7.
Mar, 2013	**Oculus Rift DK1** - developer kit of a virtual-reality headset, used for engagement-dependent interactions with virtual media in chapter 9.
Mar, 2014	**Android Wear** - wearable devices, such as smartwatches and Google Glass, become widely supported, offering compelling use cases for Query by Tapping.
Jan, 2015	**BeoSound Moment** - Bang & Olufsen device launched, building on the engagement-dependent approach and concept. The product and an exploratory evaluation is described in chapter 8.
Apr, 2015	**Thesis Completion**

Figure 1.1: Timeline of technological developments that occurred around this work.

Part I

Background

2. Music Information Retrieval

The retrieval of music from large media collections is a well-studied area. The earliest reference to Music Information Retrieval (MIR) was Kassler's (1966) paper 'Toward Musical Information Retrieval', presenting a system to assist musicological analyses. Until the 1990s, the field was largely concerned with the retrieval of detailed representations of music from small collections by musicologists. Downie (2003) terms this early branch of research as 'Analytic/Production MIR' in contrast to 'Locating MIR', the latter concerning the use of incomplete musical representations to locate a piece of music.

The recent focus in the field has been on producing such representations directly from audio sources, allowing large media collections to be indexed. Locating MIR systems have now become desirable for non-expert music listeners with access to large media collections. The field of MIR has grown to span many disciplines, from sociological studies of music use through to machine learning approaches to classifying, transcribing and even generating music. Genre classification became a large focus for many years, with advances in classifier performance approaching an apparent upper limit (Aucouturier and Pachet, 2004). This limit is posited to be due to the noisy ground truth of genre labels, and the subjective nature of genre (Sturm, 2014a). In response to this subjectivity, this chapter identifies those aspects of music which can be said to be universal. Recent developments in MIR have turned to focus on user-centred approaches to designing, adapting and evaluating music retrieval systems.

2.1 CHALLENGES

In his early overview of Music Information Retrieval, Downie (2003) identifies a range of challenges faced in the development of music retrieval systems. The first of these is the *multifaceted* challenge - the need to consider the many facets of music information, of which seven are listed. The facets given describe not only the acoustic properties of music, but also the metadata of music, such as performing artist or year of release. This difficulty in dealing with the range of music information is compounded by the *multirepresentational* challenge – that music information can be expressed textually, symbolically as a score or MIDI file, or as an audio representation, which may be recorded live or in a studio and in a variety of formats (WAV, MP3 etc.). Users' queries for music can thus be expressed in a range of representations, such as examples of music (recorded, sung or hummed), genre and mood categories, textual descriptions, or other metadata. Considering all this variety has led to the *multidisciplinarity* challenge, with researchers involved in the MIR community spanning from musicologists considering musical scores to engineers employing techniques from audio signal processing and machine learning. The challenges listed thus far only relate to the complexity involved in considering representations of music. The variability in users' interpretations of music introduces the *multicultural* and *multiexperiential* challenges. These challenges are dealt with next in this chapter, going some way towards dispelling the notion that the issue of too-much-choice can be solved simply by users handing control over to music classification and recommender systems.

MULTICULTURAL CHALLENGE

The way in which music is defined, produced and retrieved is culturally dependent. Efforts to define music have identified few properties which are universal across cultures (Higgins, 2012) (these properties are explored in section 2.6). It is unsurprising then that the musical features used by users to express music queries also vary with culture (Lee et al., 2005).

Serra (2011) launched the European research project 'CompMusic', seeking to develop MIR technologies for musical cultures outwith the western-centric focus of MIR to date. This project led to a great deal of work addressing a variety of musical cultures, from analysing Turkish Makam music (Bozkurt et al., 2014) to Jingju (Beijing opera) (Repetto and Serra, 2014). While these efforts go some way towards addressing the multicultural challenge, that this research work has had to occur in parallel for each culture only serves to emphasise the scale of the challenge. The need to consider individual cultures emphasises the moral hazard discussed in the introduction to this thesis – with the promotion of a single set of 'objective' music classifiers bordering on being an act of cultural vandalism.

MULTIEXPERIENTIAL CHALLENGE

The interpretation and appreciation of music is highly subjective, varying not only with each individual but also across that individual's listening contexts. Downie (2003) terms this the multiexperiential challenge, i.e. that 'Music ultimately exists in the mind of its perceiver.' Downie argues that 'to ignore the experiential aspect of the music retrieval process is to diminish the very core of the MIR endeavor,' concluding that the experiential nature of similarity and relevance is the most important challenge facing Music Information Retrieval. Wiggins (2009) criticises the prevailing approach in MIR of audio-only approaches with assumed ground truth. He makes a detailed argument for the study of music only in the context of models of subjective human perception, and in doing so notes that 'in MIR there is no ground.' This realisation in MIR reflects earlier developments in musicology, Leman et al. (2008) describe the trend towards a 'new musicology' which 'like postmodern thinking, assumes that there is no absolute truth to be known.'

It is uncontroversial to suggest that music recommendation should account for subjectivity, however even seemingly objective systems-centric work, such as the development of music classifiers, must consider the user-specific nature of MIR. Work continued for many years treating aspects of music such as genre and similarity as objective, at least for the purposes of this early academic investigation. The results of this work have already run up against the upper bound imposed by limited human consensus (Flexer, 2014; Sturm, 2014a). This point is explored in the following sections, as it is important to emphasise the inherently user-centric nature of music retrieval and the limitations of approaches ignoring this. The use of labels, taxonomies, rankings or distance metrics with an assumed objective ground truth is fundamentally flawed and suffers the limitations of an ill-posed learning problem.

User-adaptive approaches to classification or recommendation would still struggle to overcome the challenges of subjectivity, as this subjectivity itself is inconsistent. Born and Hesmondhalgh (2000) (p.33) explore the music identities that characterise the music listener, explaining that 'Rather than musical subjectivity being fixed and unitary, several musical 'identities' may inhabit the same individual. These are expressed in different musical tastes and practices, some of them in tension with each other or in contradiction with other parts of the self.' Given that music listeners have (perhaps multiple) subjective tastes and interpretations of music, which are coupled with cultural context, no recommender system could hope to always serve the correct choice without first modeling the entire user. Recommendation is still useful however, and the issues of individuality and subjectivity only serve to emphasise the need to offer users the choice to influence and control their music retrieval. The rest of this chapter explores the limitations that subjectivity imposes upon attempts at objective classification, and considers calls for user-centred MIR, showing that simply building better recommender systems is not the solution to the challenges addressed in this thesis.

2.2 GENRE

When discussing or searching for music, it is common to refer to its *genre* – a categorisation of the music (Lee and Downie, 2004). Rather than being part of a well-defined taxonomy, Fabbri (1982) describes musical genre as being a dynamic 'set of socially accepted rules', conditioned upon time and cultural context. Many music retrieval systems employ genre as a means for filtering or recommending music, and the automatic classification of musical genre has been a significant focus of MIR research. The motivation for such work is clear; by classifying music into genres and sub-genres, music retrieval systems can support users in accessing the volumes of music available in digital collections.

Musical works may relate to a number of genres and deviate from their norms, and Fabbri seeks to 'criticise the widespread influence of Aristotelian and positivist traditions.' It is notable then, that the majority of work in music classification assumes absolute ground-truth labels of musical genre. Given the prevalence of music classification as an approach to addressing the issue of too-much-choice, it is worthwhile to consider the performance of such classifiers and their limitations.

CLASSIFICATION

One of the key early results in music classification, by Tzanetakis and Cook (2002), achieved 61% agreement with ten genre categories using timbral, rhythmic and pitch features. They acknowledge the 'fuzzy nature of genre boundaries' however argue that there is still utility in genre classification for MIR, pointing to better than chance performance. The *GTZAN* dataset made available by Tzanetakis and Cook became a benchmark for the field, with performance increasing over the years to 93.7% reported by Panagakis and Kotropoulos (2010). In realistic usage scenarios, classifiers must support more than the 10 broad genres often considered in the literature. In a classification task involving 19 industry genre labels, Seyerlehner et al. (2010) tested a range of classifiers, with the best achieving only 45% accuracy.

The impressive performance figures reported in some of the work on music classification suggest that the problem may be all but solved. Bergstra et al. (2006) go so far as to question if 'automatic methods are not already more efficient at learning genres than some people.' Such a suggestion is not misplaced, human performance in classifying music into ten genre labels has been found to be around 70% (Gjerdingen and Perrott, 2008), even with a relatively homogeneous sample of participants (undergraduate students, mean age of 19.5). A direct comparison between classifiers and human performance by Seyerlehner et al. (2010) used a broader demographic and 19 genre labels, showing participant labelling accuracy ranging from 26%-71% agreement with genre labels.

LIMITATIONS & CRITICISMS

If ground-truth genre labels are not reliably reproducible by listeners, there is some question as to their validity – to what extent are these labels 'a set of socially accepted rules' rather than a taxonomy specific to a particular dataset. Sturm (2014a) further questions the validity of progress in genre classification, demonstrating that at least some state of the art classifiers are actually discriminating based on confounding factors and not genre. An example of such a confound would be the equalisation of a track. Given the use of spectral features in classification, it is plausible that tracks are effectively labeled by the production processes and equalisation curves specific to each genre. This raises issues of generalisability and the moral hazard regarding independently produced tracks, which may be incorrectly classified.

An analysis of the *GTZAN* dataset revealed that tracks in the dataset are almost all distorted, 5% are duplicated and 10% are mislabeled, leading Sturm (2012) to question 'the extent to which we should believe any conclusions drawn from the results.' Craft et al. (2007) note that where subjectivity is apparent, users tend to form self-similar groupings and so clustering users may address some issues with genre classification. While machine listening approaches play a crucial role in music retrieval, it is important to acknowledge the subjectivity and uncertainty involved when incorporating them into music retrieval systems.

2.3 USER-CENTRIC APPROACHES

With a large volume of user listening histories, collaborative filtering (CF) offers a way to recommend related music without having to define arbitrary genres or moods. While the approach is inherently user-centred, it has been shown to be highly biased towards popular music. Celma and Cano (2008) compare the ability of CF to retrieve music from the *long tail* against machine listening and human expert approaches, showing that CF, and to a lesser extent humans, are biased towards popularity.

The inherent subjectivity and user-centred nature of music has often been acknowledged within the field of music information retrieval however only recently have researchers begun to address it. Hu and Liu (2010) set out the case for a shift to user-centred evaluation in the field of MIR, pointing to similar arguments being made in textual IR for tasks outside of the information-seeking 'Cranfield paradigm' (Robertson, 2008). They propose evaluative measures such as usefulness, learnability, task accomplishment and task duration. In their chapter on User-Aware Music Retrieval, Schedl et al. (2012) review this new direction in MIR and set out some of the factors they believe will be key to future work, such as similarity, serendipity and transparency. They also highlight the need for linking music features to user perception. Section 5.2 thus takes a user centered approach to music feature selection, and some of the universal aspects of perception are discussed in section 2.6.

ADAPTATION

Schedl and Flexer (2012) argue for taking an adaptive approach to genre - 'the coarse and ambiguous concept of genre should either be treated in a personalized way or replaced by the concept of similarity.' Subsequent work by Flexer shows how even similarity would fall foul of the multiexperiential challenge. Flexer (2014) investigated inter-rater agreement on music similarity, finding that a lack of human agreement on whether musical works are similar places an upper bound on similarity-based approaches. Wolff and Weyde (2014) explore the benefit of adapting music similarity measures to ratings elicited from users, they also note the multiexperiential issue and propose to model cultural factors in future work. Stober (2011) presented interactive approaches to user-adaptive genre classification, music similarity and visualisation through re-weighting classification features. The need for user-adaptive MIR was also identified by Shao (2011) in his thesis, which aspires to the type of system developed by Stober. The need for user-centric, adaptive approaches to MIR is now well-established, though the practice itself has yet to be widely adopted. Further progress in this area will require a greater understanding of music listeners.

MUSIC RETRIEVAL BEHAVIOURS

User studies have provided insights about user behaviour in retrieving and listening to music and highlighted the lack of consideration in MIR about actual user needs. In 2003, Cunningham et al. (2003) bemoaned that development of music retrieval systems relied on 'anecdotal evidence of user needs, intuitive feelings for user information seeking behavior, and a priori assumptions of typical usage scenarios.' While there have been many user studies conducted since, the situation has been slow to improve. A review conducted a decade later by Lee and Cunningham (2012) noted that approaches to music retrieval system design and evaluation still largely ignore the findings of these user studies. Such a state of affairs is unfortunate, as the findings of these studies raise important questions and illuminate interesting areas for future MIR research.

In a survey comparing music retrieval requests on Korean and Western websites, Lee et al. (2005) identified a sizeable disparity in the features of music used to express a query. Where musical genre is used in half of all western queries, it was used in only 16% of Korean queries. The intended use of music was a much more significant feature in the Korean music queries. While examining how users organise their music collections, Cunningham et al. (2004) also identified the use of 'idiosyncratic genres', i.e. that music was characterised by intended use. Their use of the term genre here is striking, as it does not refer to the style of music as often assumed in MIR, or capture a 'set of socially accepted rules' as stated in musicology, however it does reflect the way music is queried across cultures.

2.4 EVALUATION

The lack of robust evaluations in the field of MIR was identified by Futrelle and Downie as early as 2003. They noted the lack of any standardised evaluations and in particular that MIR research commonly had an 'emphasis on basic research over application to, and involvement with, users.' In an effort to address these failings, the TREC-inspired Music Information Retrieval Evaluation Exchange (MIREX) was established (Downie, 2006). MIREX provides a standardised framework of evaluation for a range of MIR problems using common metrics and datasets, and acts as the benchmark for the field. While the focus on this benchmark has done a great deal towards the standardisation of evaluations, it has distracted research from evaluations of complete music retrieval systems with real users in realistic contexts.

LIMITATIONS OF EXISTING APPROACHES

A large amount of evaluative work in MIR focuses on the performance of classifiers, typically of mood or genre classes. A thorough treatment of the typical approaches to evaluation and their shortcomings is given by Sturm (2014a,b). Virtually all such evaluations seek to circumvent involving users, instead relying on an assumed ground truth, e.g. the widely used ground truth dataset *GTZAN*, a small collection of music with the author's genre annotations. Even were the objectivity of such annotations to be assumed, such datasets can be subject to confounding factors such as equalisation, and mislabellings as shown by Sturm (2012). Schedl et al. (2013) observe that MIREX evaluations similarly involve assessors' own subjective annotations as ground truth. They criticise the systems-centric evaluations that 'completely ignore user context and user properties, even though they clearly influence the result.' In his work on inter-rater agreement in music similarity, Flexer (2014) concludes that 'any evaluation of MIR systems that is based on ground truth annotated by humans has the same fundamental problem.' The implications of this issue are far-reaching, the assumption of an objective ground truth for music genre, mood etc. is common (Craft et al., 2007), with evaluations focusing on these rather than considering users.

USER-CENTRED EVALUATION

There remains a need for robust, standardised evaluations featuring actual users of MIR systems, with growing calls for a more user-centric approach. Schedl and Flexer (2012) made the broad case for 'putting the user in the center of music information retrieval', concerning not only user-centred development but also the need for evaluative experiments which control independent variables that may affect dependent variables. There is, in particular, a need for quantitative dependent variables for user-centred evaluations.

For well defined tasks such as audio similarity or genre classification, existing dependent variables may be sufficient, though Sturm (2014a) criticises typical genre classification evaluations. If, however, the field of MIR is to concern itself with the development of complete music retrieval systems, their interfaces, interaction techniques, and the needs of a variety of users, then new evaluative approaches and additional metrics are required. The need for user centered evaluations in MIR has been acknowledged within MIREX, with the introduction of the 'Grand Challenge on User Experience' (GCUX), involving ratings of complete systems by users. Given that the purpose of a Music Retrieval system is to support the user's retrieval of music, a dependent variable to measure this behaviour is desirable. Such a measure cannot be acquired independently of users – the definition of musical relevance is itself subjective. The measures of music listening and organisation introduced in part II of this thesis aim to quantify the exploration, diversity and underlying mental models of users' music retrieval.

2.5 QUERYING METHODS

A wide range of querying techniques have been explored for music, generally falling into two categories: content-based and those based upon more abstract representations such as keywords. Content based techniques involve the use of acoustic data such as a recording or melody to find matching music and are thus more 'low-level' approaches. Other techniques such as keyword search or the navigation of a hierarchy of categories are 'high-level' in approach, requiring the use of machine learning, expert labels or collaborative filtering.

In their overview of content-based MIR, Casey et al. (2008) illustrate a general model for a 'canonical content-based query system' in which queries of content are matched to audio tracks. Examples of such a system are Shazam[1] and Soundhound,[2] where music can be queried by a recording or user rendition respectively. Shazam employs audio 'fingerprinting' or spectogram hashing (Wang et al., 2003), and can robustly match recorded audio to the exact same recording in a database. SoundHound instead supports users reproducing music in order to query for it, building upon a body of work in MIR on *Query by Humming* (Ghias et al., 1995). Both systems achieve high levels of performance, and have thus become very popular with users. Content-based methods are relatively free from issues of subjectivity, with the content itself providing the objective ground truth, however a similarity function must still be defined. The recording of audio, and especially the performance of a musical piece, is however a querying style not available in many contexts, requiring a high degree of user engagement (and lack of a bashful nature).

[1] http://thenextweb.com/insider/2014/08/20/shazam-now-100-million-monthly-active-users-mobile/

[2] http://www1.soundhound.com/index.php?action=s.press_release&pr=67 (06/12/14)

2.6 MUSICAL UNIVERSALS

Unsurprisingly, there are a great many differences in music throughout the world – to the extent that a complete definition of what music is remains elusive. Higgins (2012) explores to what degree music can be defined and what aspects of music are universal across cultures. Remarkably little can be taken for granted, with even the emotional valence response to music being culturally dependent (Egermann et al., 2015). Some aspects of music that can be considered as being universal are discussed below, with a view to their use in developing music information retrieval systems that are robust against the multicultural challenge.

MELODIC CONTOURS

It was pointed out by Helmholtz (1895) that melodies are made up of discrete steps in pitch, purported by him to be due to the need for greater cultivation of ear to distinguish finer details. Dowling (1978) identifies melody as comprising two components, musical scale and melodic contour as a 'pattern of ups and downs' in pitch. It was shown that melodic contours are cognised independently of musical scale and are overlaid upon the scale with which a listener is familiar. Indeed, Perlman and Krumhansl (1996) showed that pitches are heard within the categories of the tuning system(s) the listener is familiar with and that this effect exists across cultures. This categorisation of pitch enables the encoding of melodic contours and is an assumption made in MIR systems such as *Query by Humming*. This categorical perception of music is not limited to pitch, with rhythm being another notable example.

RHYTHM

Rhythmic patterns are fundamental to music, with work by Trehub (2000) showing that rhythmic patterns are more important for cognising musical sequences than absolute position of events in the time domain. Surprisingly, Monahan and Carterette (1985) showed that rhythm is a greater factor for people when they assess the similarity of musical patterns than pitch. The use of rhythm then is perhaps the foremost musical universal. Crucially, the way in which people process rhythm has been proposed by Drake and Bertrand (2001) to be universal across cultures. They propose as a universal musical principle that 'We tend to hear a time interval as twice as long or short as previous intervals.' Complex rhythms are thus distorted into categories of intervals between note onsets (*inter-onset intervals*); this principle is exploited in chapter 6 where rhythm is encoded in terms of these categories.

3. ENGAGEMENT & CHOICE

ADDRESSING the issue of too-much-choice in music retrieval requires the consideration of a number of intersecting fields. First and foremost are recent developments in music psychology, which have begun to elucidate the differing relationships listeners have with music in terms of their *music engagement*. The extent to which listeners wish to control and take ownership of their music selection is captured in this conception of engagement, and early work has produced some questionnaires to measure it. Similar efforts in the field of Interactive Information Retrieval have considered retrieval scenarios where users are not seeking to complete a defined task, but often to casually browse with a reduced level of effort. A further perspective considered is that of the user's *interaction engagement* – how much the user provides input and attention to the retrieval process, and the extent to which they seek to *control* it. These lines of inquiry are thematically linked in discussing the amount of effort and information that a user is willing to provide to make a satisfactory music selection. This chapter explores this theme, with further empirical illumination given in chapter 4.

3.1 TOO MUCH CHOICE

The growth in the size of music collections and the availability of music online has dramatically increased the choice facing users when selecting music. While increased choice has been empirically shown to improve motivation and satisfaction, a growing body of literature claims that too much choice can overload users, in what is termed the 'too–much–choice' effect (Iyengar and Lepper, 2000; Schwartz, 2005; Scheibehenne et al., 2009). Requiring users to make an explicit selection from many choices incurs greater cognitive load and selection time (Hick, 1952; Hyman, 1953).

The problem of too much choice applies when the range of options becomes unmanageable, leaving the user paralysed and indecisive. Faced with the burden of making a selection from many options, users may become demotivated, opting instead to shuffle music, listen to the same few artists, or even turn on the TV for sound instead. Scheibehenne et al. point out that users may be unhappy with making a choice when there is a subsequent requirement to justify these choices. Given that music selection is often part of a social interaction, and that a user is judged by their selections (Cunningham et al., 2004), this effect will likely apply. Music listening also often occurs in mobile contexts, such as when walking or driving, where users are limited in the attention they can dedicate to making choices.

SATISFICING & SELECTION

In a study of the music-seeking behaviour of young adults, Laplante (2008) observed that music listeners often have no particular song in mind when engaging with a music system, and instead browse for a suitable piece of music. This activity of finding the first 'good enough' item is termed *satisficing* whereas trying to optimise to find the best option is termed *maximising*. Schwartz et al. (2002) explored maximising versus satisficing as a personality construct, claiming that individuals vary with respect to how much they optimise selections. The range from satisficing to maximising can be considered in terms of information need – the more users maximise, the more well-defined their information need and the smaller the set of relevant documents (or tracks of music).

The satisficing-maximising construct captures the range of music retrieval behaviours – where users do not wish to expend any effort, they satisfice by using random shuffling or recommendation, seeking a serendipitous music selection (Leong et al., 2005). Whether a user wishes to satisfice or make more exact selections also depends upon their present context and how much they wish to engage in their music retrieval. The extent to which the too-much-choice effect applies is thus user and context-dependent, and the appropriate level of filtering and recommendation versus user control will change accordingly.

3.2 MUSIC ENGAGEMENT

The field of Music Psychology has explored the concept of *music engagement*, given by Greasley (2008) as:

> The extent to which music plays an integral role in a person's everyday life, including the importance of music; the amount of music owned; uses of music; involvement in music-related activities; and various other factors ... people's levels of engagement with music represents a spectrum, with clearly identifiable extremes (e.g. those who are either very highly engaged or who rarely engage with music in daily life), and overlaps in the middle (e.g. those showing characteristics of both high and low engagement.

It is clear that this concept of engagement is of importance to music retrieval, however this definition refers to the relationship between listener and music, rather than the user's engagement in a specific interaction or act of retrieval as considered in this thesis. Greasley goes on to link music engagement to a user's desire to control the music they listen to, with lower levels of engagement linked to radio usage rather than controlled song selection. Work within the field of Music Information Retrieval has also touched on the idea of users having differing degrees of involvement with music, Celma (2010) illustrates the range from *musical indifferents* to *musical savants* as captured in the Phoenix 2 project.

MEASUREMENT

As a relatively new area of research, there has been little work exploring how to measure a user's music engagement. Greasley (2008) used 5 point Likert scales for a series of largely qualitative questions, aimed at different factors of music engagement elicited in user interviews. These factors were:

Control The user's preferred level of control over the selection of music that they listen to.

Collection The size of a user's music collection.

Organisation The extent to which the user organises their music collection.

Motivation A measure of how motivated the user is in acquiring music for their collection.

Memory The detail of a user's memory of their first music purchase.

Lyrics How important the user considers lyrics to be when listening to music.

The use of experience sampling with wider Likert scales for measuring music engagement was introduced by Greasley and Lamont (2011), demonstrating that more engaged music listeners tended to have greater control over the music they heard, often selecting it to evoke specific moods. Related work by Krause et al. (2014) further explored how users control their music selection, also using experience sampling methodology (ESM) to show that music selection behaviour changes with music engagement as well as age. The use of ESM allows for an in-depth analysis of a user's music-listening behaviour, however it is time-consuming to run and difficult to scale to many participants. Notably, ESM illuminates the variance in behaviour between sessions of music-listening within one participant – music engagement varies with listening context, including situational, social and multi-tasking factors as well being an individual characteristic (Greasley and Lamont, 2011). The use of these questionnaires and their relationship to other measures of engagement is explored in section 4.3.

3.3 INTERACTIVE INFORMATION RETRIEVAL

The process of a user making a choice from a collection of objects is considered in Interactive Information Retrieval (IIR), with efforts to model users' engagement in the retrieval process. Of particular relevance to this thesis is what Wilson and Elsweiler (2010) describe as *casual-leisure searching*, where users do not have a focused goal of satisfying an information need but are instead interacting for hedonic purposes. They point out that these casual search scenarios are ones that 'break our current models', and identify behaviours such as 'needless' browsing of TV channels. This area of research is highly related to the concept of engagement discussed in this thesis, reflecting the specificity of user queries and the amount of effort and attention that users wish to invest in their retrieval.

There is often an assumption in IR that users are focused on satisfying some information need, applying as much effort and concentration as required to maximise the gain of the retrieval. Robertson (2008) describes the standard 'Cranfield' approach, which 'is to reduce the user variables to requests and relevance judgements.' He notes how useful this abstraction has been but underlines that 'it does not allow us to answer all the research questions we might reasonably ask.' Such an approach allows for users to be modelled in systems-centric evaluations however the effort required of a user (and that a user is willing to invest) is a key aspect of designing interactive retrieval systems. In discussing evaluation in IR research, Sakai and Dou (2013) state the need to balance systems-centric evaluations with user-centric evaluations, noting that 'most metrics do not consider the user's actual effort for finding information.' They voice a question on behalf of user-oriented researchers in the field – 'Where's the user?'

THE COST OF CHOICE

Azzopardi (2011, 2014) introduced an economic model of search behaviour, showing how the cost of interacting with a retrieval system changes a user's search behaviour. While this model was developed with textual retrieval in mind, it is of value to consider music retrieval in terms of interaction costs – in particular, how the costs of interacting with a retrieval system vary according to user engagement and context (e.g. querying textually on a mobile device requires more effort). In later work, Jiang et al. (2015) used Azzopardi's model to establish a link between the user's effort in a search and their search satisfaction. The results showed that user satisfaction can be considered as 'search outcome compared with search effort' – users adjust their expectations according to their level of effort, and are satisfied if the system meets these expectations. It is worth noting that the search satisfaction ratings used by Jiang et al. were provided by 'third-party judges' rather than the original users issuing queries. While this is a significant limitation, it would be difficult to motivate users to rate their searches. They go on to observe that a user's level of effort varies between search sessions, even though the search interface (Bing)[1] remained the same.

The cost of too-much-choice has been considered in many fields, beyond IR or the work of Schwartz (2005). Modelling how humans make choices with limited cognitive and temporal resources is very much an ongoing effort. Simon (1990) terms this behavioural modelling of choice *bounded rationality* and contrasts this with models that consider people's choices as the maximisation of a utility function and that do not account for the limited reasoning power and knowledge available.

3.4 INTERACTION ENGAGEMENT

Consideration of user engagement has covered not only the user's involvement in an interaction, but also a number of affective aspects, such as the user's motivation and emotional investment in the outcome of the interaction. A single definition of user engagement would thus be over-broad and Lalmas et al. (2014) describe their effort at a comprehensive definition as 'clunky'. In earlier work, O'Brien and Toms (2008) listed attributes of user engagement: 'challenge, affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, perceived control, and interactivity.' While it is difficult to quantify attributes such as affect and aesthetic appeal, some of the attributes can be grouped into a more tractable concept. The feedback and interactivity of a system are easily quantifiable, and will directly relate to the user's perceived control and attention. This control element of engagement is of interest, as it can be related to user control as discussed in music engagement, and later in this section in discussing control-theoretic views of interaction.

[1] `http://www.bing.com` (16/12/14)

MODELS OF ENGAGEMENT

Lehmann et al. (2012) argue that, given the complexity of the concept of user engagement, evaluators should adopt models, not metrics, of engagement. While time on task, dwell time, user clicks and other simple measures may seem attractive dependent variables for evaluation and experimentation, the effects seen require a great deal of interpretation. For example, increased time on task could indicate an inefficient retrieval interface, or alternately it could be the result of a user engaging with a compelling interface that afforded exploration. Dupret and Lalmas (2013) proposed absence time as a measure of engagement, hypothesising that engaged users will be more likely to regularly return to a service. They adopted Survival Analysis techniques to model this behaviour, treating users' return visits to a service as comparable to patient deaths as typically modelled with survival and hazard functions. This approach is similar to the work in chapter 4, which considers how far a user 'survives' into a track before deciding to skip. The use of survival analysis has also recently been introduced to the music retrieval field, in work by Kapoor et al. (2014) funded by Pandora. They use Last.fm and proprietary datasets of user listening history to classify users by their predicted return time, identifying users with low engagement that a business should re-engage with.

FOCUS & CONTROL

O'Brien and Toms (2010) developed a survey to capture and explore attributes of user engagement, and evaluated them in two large studies. While their findings relate to online shopping, they give an illustration of the relationships between the many attributes associated with engagement. A principal components analysis of their survey identified six factors of engagement, which they labelled as: Focused attention, Perceived usability, Aesthetics, Endurability, Novelty, and Involvement. Of these, focused attention and perceived usability were the most significant, accounting for 29.73% and 15.63% of the variance respectively. Focused attention included questions regarding absorption and awareness of time passing, concepts related to Flow theory (Csikszentmihalyi et al., 2014). Perceived usability included questions regarding control, effort, users' ability to accomplish their tasks and the emotions evoked. The *focus* factor captures the user's attention and perception of the interaction and the perceived usability factor captures the user's ability to *control* the interaction.

In their work on user engagement in interaction, Pohl and Murray-Smith (2013) use control-theory to describe users' control over an interaction. Their work goes some way towards a robust characterisation of engagement, identifying similar concepts to O'Brien and Toms (2010). It is notable that amongst the various considerations of engagement, the user's level of control is a consistent factor. While each definition of engagement encompasses its own mixture of factors, a common factor is the user's control over the interaction. This control factor is also more readily quantifiable than the affective aspects of engagement.

In Pohl and Murray-Smith's control-theoretic viewpoint, user engagement is the degree to which the user is in a control loop with the system, directing it towards some goal. This thesis considers the retrieval of music specifically, and the goal-directed retrieval behaviour (and the extent to which the user is controlling a system to satisfy an information need) is referred to as the *retrieval control*. This retrieval control is a generalisation of the control factor of music engagement. In Pohl and Murray-Smith's description, user engagement is the user investing more attention and effort into an increasingly precise and tightly-coupled interaction loop with the system. Wilson and Elsweiler (2010) noted that casual scenarios break current models in IR, and by considering the user's desired retrieval control in a casual scenario, some progress can be made to remedy this issue. These casual interaction scenarios can be modelled by linking casual retrieval tasks to the uncertainty about the user's goal and reduced control from the user. Challenges remain in exactly how to model the uncertainty of the user's goal (e.g. the objective function in a satisficing task) and developing a more detailed model of how the user's *focus* in a retrieval interaction relates to their *retrieval control*.

RETRIEVAL CONTROL

The theme of user control and in particular *retrieval control* is of greatest interest in considering Information Retrieval scenarios, as the user's efforts in reaching a goal can be related to the specificity of their Information Need, and how much effort they will invest to satisfy that need. In their description of how an interactive system could support a range of user engagement (control) levels, Pohl and Murray-Smith (2013) invoke the 'H-metaphor' from Flemisch et al. (2003), where a horse responds to the rider's control via the reins, however acts more autonomously where the rider does not assert control. A retrieval system could similarly allow users to vary their degree of control, increasingly relying on autonomous recommendation as the user provides less authoritative input.

Pohl and Murray-Smith's view of user engagement as being an increasingly tight coupling between user and system thus provides an insight into how music retrieval systems can act autonomously, only insofar as the user desires. Crucial to this approach is the ability to correctly determine the user's desired engagement. Despite limited apparent engagement, power users may provide the occasional precise, controlling input, such as the album-based listening identified in chapter 4. Similarly, users who satisfice and have a minimal *retrieval control* may nonetheless have a high degree of engagement in the played music. This *media engagement* – the user's attention and focus to the retrieved media – may even at times compete for the same cognitive resources as the user's *interaction engagement* with the retrieval system. This scenario is explored in the extreme case of immersive media, such as Virtual Reality, in chapter 9 and highlights the importance of Lehmann et al.'s call for models of engagement to capture the complex dynamics of user behaviour.

3.5 INFORMATION-THEORETIC VIEW OF ENGAGEMENT

Information theory provides a framework for discussing the uncertainty of the user's goals. In particular, it provides a means of quantifying how much the user's desired *retrieval control* reduces the uncertainty about their intended music selection. Interaction is defined in ISO 9421 as the exchange of information between a user and their computer. The information communicated in an interaction can be measured in *bits*. In an interaction, there is an uncertainty about the user's intent and by providing bits of information the user reduces this uncertainty, bringing the interaction towards their intended state. Empowerment is the upper limit to an agent's (or user's) communication and influence, or more formally, 'the channel capacity of the agent's actuation channel' and perception channel (Klyubin et al., 2005). This rate of influence and information exchange can be measured in *bits per second* and is the maximum bandwidth for an interaction. Engagement could then be considered as an agent's degree of participation in an interaction, i.e. to what extent the user wishes to communicate information with their computer (so as to bring it towards a desired state).

It follows from the above definitions that engagement determines the *desired* channel capacity through which a user communicates and consumes information. More informally, it is the amount of cognitive and physical bandwidth the user is willing to utilise to achieve their goal. This is analogous to but distinct from the channel capacity of the interactive devices used, which may act as a bottleneck for the interaction or be under-utilised. This information-theoretic view of engagement is somewhat reductive compared to the various definitions reviewed in this chapter, which extend to affective involvement. It does, however, reflect the use of information theory in early Human Computer Interaction (HCI) work such as Fitts' law (1954) and provides a theoretical framework in which user engagement can be linked to other fields such as information retrieval, control theory, and empowerment.

As music selections are made by either the user, or a recommender system, or both in part, the control over the interaction can be considered in terms of either of these agents. The more an agent provides information to reduce the uncertainty about the music to be played, the more control that agent has exerted. *Retrieval control* can thus be measured quantitatively, as the difference between the entropy over a music collection and the entropy over the selection made. A shuffle-based music retrieval system would not reduce this uncertainty, providing 0 bits of influence. A recommender system would likely reduce the uncertainty significantly, and the user making a specific album selection would reduce nearly all uncertainty over the music collection. Much of the information communicated in an interaction may not directly reduce the uncertainty of the selection, as the user may attend for affective or exploratory reasons.

3.6 SYNTHESIS

The various conceptualisations of engagement are used throughout this thesis. For the sake of clarity, they are differentiated explicitly as:

Music engagement is a measure adopted from music psychology, capturing the importance of music and control over music selection to a listener.

Interaction engagement relates to how much of the user's focus and perceptual attention is invested in an interaction, though may also span to affective aspects such as emotional involvement in the interaction.

Media engagement relates to how much of the user's focus and perceptual attention is invested in the consumption of media. This can, at times, be in competition with their interaction engagement, e.g. when immersed in a VR environment (chapter 9), or when producing a rhythmic query (as detailed in chapter 6) while listening to music.

Retrieval control is introduced as a robust definition of the control factor of music engagement, being how much the user reduces the uncertainty of playback intent over a music collection.

It is noteworthy that research efforts considering engagement in music psychology, Interactive Information Retrieval and Human Computer Interaction all touch upon the common theme of control. Considering the user's *retrieval control*, and their provision of information to reduce the entropy of music playback, is useful in addressing the problem of 'too-much-choice'. Interactions often offer a fixed degree of *retrieval control*, with a large amount of information required to reduce the uncertainty about what music to play, e.g. a textual search box or a hierarchical artist-album-track menu. Where the expected input of information (the specificity of the query) is inappropriate for the user's context, the user suffers from the burden of control and too-much-choice. This thesis thus deals mainly with how a user's desired *retrieval control* shapes their music listening behaviour (part II) and how music retrieval interfaces can be designed to support these behaviours (part III).

While the attention a user invests in an interaction is likely to relate to their desired control, this may not always be the case. The example of immersive media is considered in part IV, where the user's *media engagement* and attention to their media experience precludes their *interaction engagement* and ability to control the interaction. Virtual Reality is considered as a topical, extreme case of this issue, where a balance must be struck between the user's immersion in their VR experience, and their interaction with modalities to control the experience.

Part II

Understanding Music Listening

4. Measures of Music Listening Behaviour

UNDERSTANDING how listeners interact with music retrieval systems is of fundamental importance to the field of Music Information Retrieval. The design and evaluation of such systems is conditioned upon assumptions about users, their listening behaviours, and their interpretation of music. While user studies have offered guidance to the field thus far, they are mostly exploratory and qualitative (Weigl and Guastavino, 2011). The availability of quantitative metrics would support the rapid evaluation and optimisation of music retrieval. This chapter develops an information-theoretic approach to measuring users' music listening behaviour, with a view to informing the development and evaluation of music retrieval systems. To demonstrate the use of these measures, a dataset 'Streamable Playlists with User Data' *(SPUD)* was compiled, comprising 19,225 playlists from Last.fm[1] produced by 7,666 users, with track metadata including audio streams from Spotify.[2] This dataset was combined with the mood and genre classification of Syntonetic's Moodagent,[3] yielding a range of intuitive music features to serve as examples.

[1] `http://www.last.fm`
[2] `http://www.spotify.com`
[3] `http://www.moodagent.com` Last accessed: 30/04/14

4.1 THE *SPUD* DATASET

The *SPUD* dataset consists of 19,225 playlists scraped[4] from the profiles of Last.fm users who were active throughout March and April, 2014. The tracks for each playlist were linked to a Spotify stream, which was used to scrape metadata such as artist, popularity, duration etc. The number of unique tracks in the dataset is 737,362 from 7,666 users. The distribution of playlist lengths is shown in Figure 4.1. The dataset is augmented with proprietary mood and genre features produced by Syntonetic's Moodagent. This is done to provide high-level and intuitive features which can be used as examples to illustrate the techniques being discussed. It is clear that many issues remain with genre and mood classification (Sturm, 2014a) and the results in this work should be interpreted with this in mind. The aim in this work is not to identify which features are best for music classification but to contribute a general approach for gaining an additional perspective on proposed music features. Another dataset of playlists *AOTM-2011* is published (McFee and Lanckriet, 2012), however the authors only give fragments of playlists where songs are also present in the Million Song Dataset (*MSD*) (Bertin-Mahieux et al., 2011). The *MSD* provides music features for a million songs but only a small fraction of songs in *AOTM-2011* were matched in *MSD*. The *SPUD* dataset is distinct in maintaining complete playlists and having time-series data of listening behaviour.

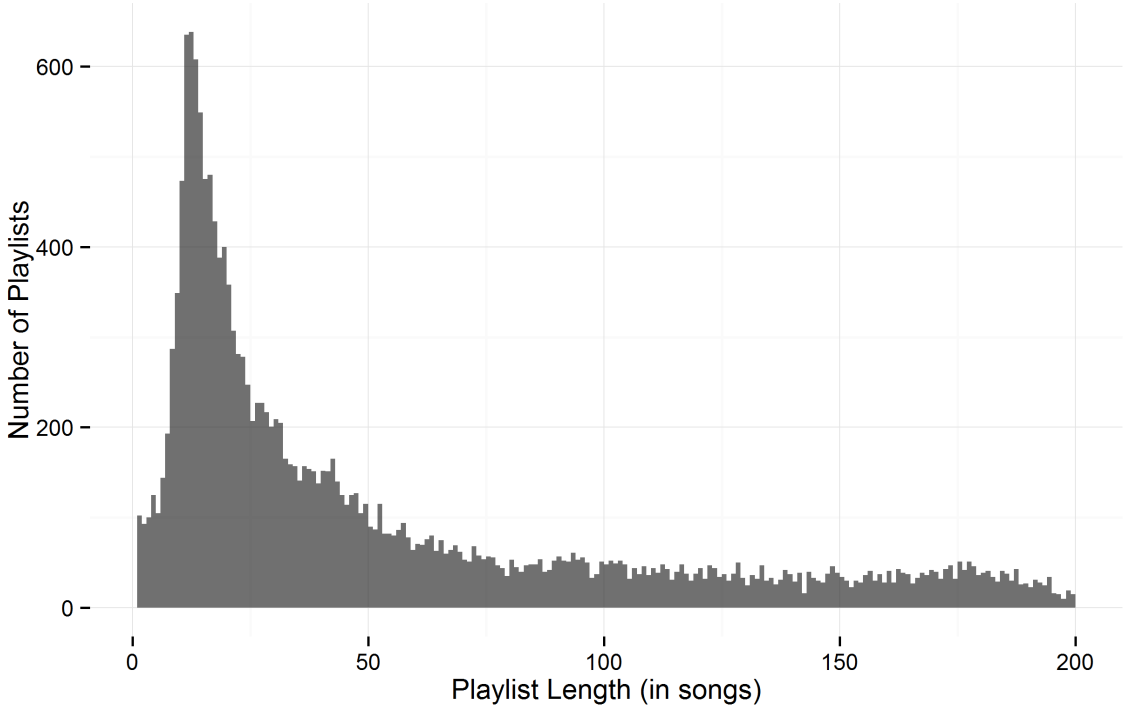

Figure 4.1: Distribution of playlist lengths in acquired dataset

[4]The playlist scraping code is available at `http://github.com/DCBoland/playlistCrawler`.

4.2 MEASURING USER INTERACTION

Listening histories were acquired from Last.fm,[5] along with the unique Spotify track IDs which were used to retrieve Spotify metadata. From these histories, it is possible to derive a number of measures about the user's interaction with their music retrieval system. Adding a song's duration to the time its playback started gives an indication of when the song would end without further intervention. Where a new song begins immediately after the expected end of a previous song, this is indicative of automatic playback such as playlists or albums Songs which are interrupted early by a subsequent song, or which are followed by another song after a brief pause, are evidence of the user intervening to find new music to listen to.

The *SPUD* dataset includes both listening histories and song durations for Last.fm users, which are used here to illustrate a number of measures and their relationships. This section explores how much of the user's music listening behaviour can be characterised in a generalised fashion, using only these interaction timestamps as opposed to the use of audio features explored later in chapter 5. These measures are defined using Iverson brackets, which map Boolean values to the integers 0 and 1 (Graham et al., 1994, p24).

ALBUM-BASED PLAYBACK

Despite the shift to digital music consumption, listening to music in the form of albums remains a common music-listening behaviour. It is useful to identify where users are listening to an album, as this provides a clear signal that the user is not selecting each individual track, or using a radio-like music retrieval. Detecting album-based selection behaviour from music listening logs is trivial, as consecutive tracks will have the same album metadata A_s. It is assumed in this work that where consecutive tracks are played from the same album, this was due to their being from the same album. The ordering of tracks on the album is not considered, as this would introduce a confound with shuffle versus sequential playback. The detection of a selection of n tracks being an *Album*-based behaviour is thus:

$$Album = \frac{1}{(n-1)} \sum_{s=2}^{n} [A_s = A_{s-1}].$$

As this value is the average of a set of Boolean values, its results are in the range $[0, 1]$. It indicates the proportion of tracks listened to that were played due to being from the same album as the preceding track. The first track is not included as it could not be based on a preceding album. Even in entirely album-based listening, subsequent albums are likely to be selected, representing some interaction by the user. As such, it is unlikely for $Album = 1$.

[5] The Last.fm scraping code is available at `http://github.com/DCBoland/listenScraper`.

INTERVENTIONS

The intervention made by a user where the song is changed before it completes playback is termed *Skipping*. Where a user hesitates after the completion of a song, in the act of manually choosing the next song, this is indicative of *Selecting* behaviour. These behaviours can be captured by comparing the timestamps logged when a track begins playing against the expected durations for those tracks. For a song s with start time t_s, song duration d_s and subsequent song start time t_{s+1}, the deviation Δd_s from the expected song playback duration is given as:

$$\Delta d_s = t_{s+1} - (t_s + d_s).$$

Considering the sign and magnitude of this deviation gives an insight into whether there were human delays or interventions, as opposed to immediate automatic playback of the next track at the end of the current one:

$$Skipping = \frac{1}{n-1} \sum_{s=1}^{n-1} \left[\Delta d_s < 0\right].$$

It is assumed that automatic song playback would occur within a threshold of $c_{auto} = 100ms$ of the previous song ending. It is also assumed that a user would intervene to switch or select a new song within a threshold of $c_{manual} = 30s$, yielding:

$$Switching = \frac{1}{n-1} \sum_{s=1}^{n-1} \left[\Delta d_s > c_{auto} \wedge \Delta d_s < c_{manual}\right].$$

The assumed value for c_{manual} is conservative to ensure that the measured pauses are part of a continuous music retrieval interaction. The aim is to detect whether a song was selected automatically or by manual intervention; delays greater than $30s$ become increasingly likely to be due to external events, for example the user pausing music playback or the end of a music listening session.

User intervention in music listening can be considered as a Bernoulli process, i.e. a binary choice between whether a user selected each song or the selection was part of automatic playback. The examples in this work consider users across all their listening sessions, with the assumption that interventions are part of a homogeneous process (with fixed probability). In practice, user behaviour will inevitably change with their listening context, however these assumptions still allow for a user's overall behaviour to be modelled. Users can thus be characterised in terms of their *Intervention Rate* – the rate at which they intervene with a *Skipping* or *Switching* event.

SPEED

While the occurrence of an intervention may provide some information about music-listening behaviour, it was hypothesised that the speed at which users intervene may add further detail. To explore this, two measures were proposed: *Switching speed* – how quickly a user switches to a new track after a previous track ends, and *Skip speed* – how far, proportionally, into a song does a user decide to skip track. The decision to normalise against the duration of the song is based on the assumption that a longer song will have a longer introductory segment, with the user needing to listen longer before deciding to skip. Considering only the j songs where a *Skipping* event has been detected:

$$Skip\ speed = \frac{1}{j-1}\sum_{s=1}^{j-1}\frac{d_s}{\Delta d_s}.$$

Switch speed can be calculated in a similar manner. Considering only the k songs where a *Switching* event has been detected:

$$Switch\ speed = \frac{1}{k-1}\sum_{s=1}^{k-1}\frac{1}{\Delta d_s}.$$

ON THE USE OF THE HARMONIC MEAN

Sometimes listeners skip near the end of a song's playback, especially where a song is fading out. The time taken to skip would be large in these instances and would bias the arithmetic mean of time taken. For this reason, the harmonic mean of skip time taken is calculated instead, mitigating the effect of large outliers:

$$Skip\ time = \frac{j-1}{\sum_{s=1}^{j-1}\frac{d_s}{\Delta d_s}}.$$

It is simpler to consider this in terms of speed by taking the reciprocal of the above. The equation then simplifies to being the arithmetic mean of speed:

$$Skip\ speed = \frac{1}{j-1}\sum_{s=1}^{j-1}\frac{d_s}{\Delta d_s}.$$

Note that when considering a rate, such as speed, one would typically use the harmonic of speed rather than arithmetic mean. The measure of interest, however, is not the average speed across all of the interaction time but is the expectation of the speed of a given selection.

Unlike the arithmetic mean of time, the harmonic mean of Δd_s is robust against slow outliers where a large time was taken (e.g. skipping as a song is nearly complete).

4.3 MEASURING MUSIC ENGAGEMENT

This section explores the utility of the music engagement questionnaires discussed earlier in section 3.2, as well as their relationship to the measures introduced in this chapter. The music engagement and music selection questionnaires from music psychology were combined and updated for digital collections (see Appendix B), and posted along with a request for Last.fm usernames to an online message board for Last.fm users. Users of Last.fm and in particular, those accessing related message boards, will be a biased sample of more musically engaged individuals. This issue is inherent in an open survey inviting participation and any interpretation of the results in this section should bear in mind that they may not generalise beyond this demographic. 95 responses were received, shown in Figure 4.2, of which 91 usable listening histories were mined. 88% reported use of mobile music selection, 93% use of a computer, 60% use of streamed music and only 18% use of radio.

The questionnaire responses were mostly skewed toward high engagement with music. The importance of *Lyrics* has a distinctly different distribution. Cronbach's α for all factors was 0.43, indicating that the questionnaire is covering a number of distinct facets of users' music engagement. How much *Control* a user likes to have over music selection is the most relevant of the factors to the measures of intervention, with the others addressing more qualitative aspects. This *Control* factor is notable for being similar to the concept of *retrieval control* as discussed in chapter 3. The measures of *Album* based listening, *Intervention Rate*, *Skip speed*, and *Select speed* were calculated from the listening histories of the respondents. Their correlations with *Control* are given in table 4.3.

LIMITATIONS & BIAS

The recruitment for this questionnaire and analysis of user listening histories was specifically targeted at Last.fm users. In order to measure a user's music listening behaviour, the user must first have logged their music listens across their devices. Performing this level of logging, and being a Last.fm user, could itself be considered a mark of music engagement. It is unlikely then that this method of recruitment will have captured the most casual of music listeners. This would explain why most of the participants who completed the questionnaire reported a high level of engagement with music in the questionnaire responses. The results explore the correlations between engagement and listening behaviour and are not intended to reflect the distribution of engagement levels amongst general music listeners.

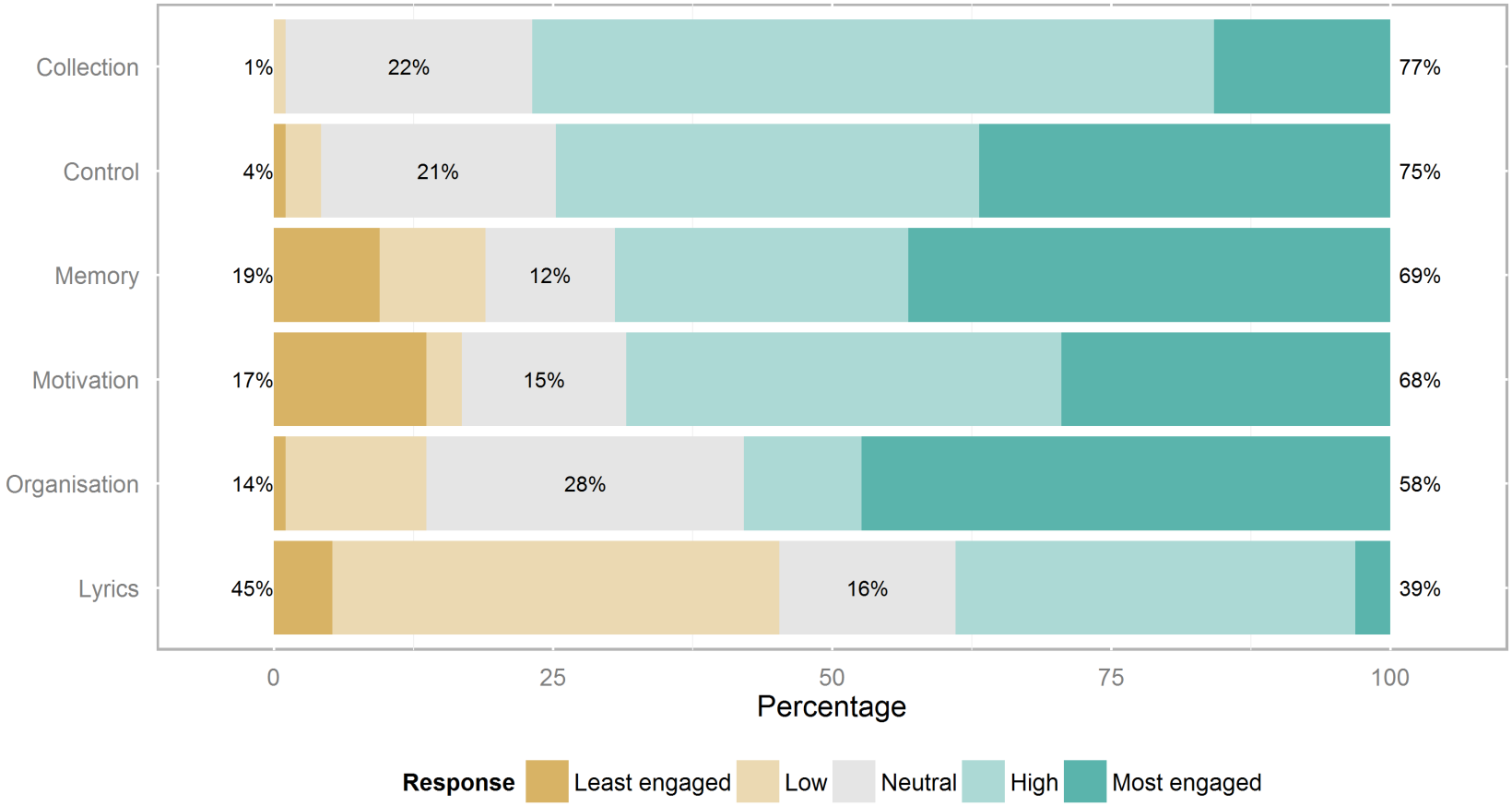

Figure 4.2: Likert-scaled music engagement questionnaire responses. Full wordings of questionnaire and available responses are included in Appendix B. Percentages are shown for aggregated low, neutral and high responses.

RESULTS

As the questionnaire responses are ordinal and the other measures are interval, polyserial correlation (Fox, 2010) was used to estimate the relationship between underlying continuous distributions from which the results were sampled. The number of tracks selected due to user *Intervention* (skipping or pausing) in music selection was significantly inversely correlated with their self-reported *Control* ($\rho = -0.29$, $p = 0.01$). Users' use of *Album* based music-listening was also significantly correlated with self-reported *Control* ($\rho = 0.26$, $p = 0.01$). The album-listening detection metric was confirmed using the questionnaire responses, with an independent-samples t-test between users who had or had not listed album-based selection, with significant differences in *Album*, *Skip speed* ($p < .01$), and *Interventions* ($p < .05$). The internal reliability of the measures in table 4.3 was high, with Cronbach's $\alpha = 0.81$.

The results in table 4.3 paint a somewhat surprising picture of how music engagement influences user interaction. While it might appear counter-intuitive that desire for *Control* leads to fewer interventions, engaged listeners are more likely to invest in selecting a specific album. This *Album* based listening was seen to have a significant moderate correlation with users' self-reported preferred level of *Control*. Naturally, album based listening also meant a significant inverse correlation with user intervention in music playback, thus the controlled selection of an album often results in less measured interaction. Casual music listening, by shuffle or recommendation, requires less initial investment, but then more corrective actions are required when songs are deemed unsuitable by the user. This reflects the nuanced view of engagement discussed in section 3.2, with *Control* proving a useful factor however not explaining the entire story of user engagement.

Control also had significant correlations with *Skip speed* and *Select speed*. More musically engaged listeners made quicker interventions in their music listening. The speeds were also negatively correlated with the number of *Interventions* i.e., the more quickly that the users intervened, the fewer interventions they made overall. These results build a profile of musically engaged listeners who know what they want. After making a selection, they make few further interventions, which are decisive and quick.

	Control	Album	Interventions	Skip Speed
Album	0.26*			
Interventions	-0.29*	-0.51***		
Skip Speed	0.48***	0.29***	-0.74***	
Selection Speed	0.25*	0.51***	-0.84***	0.47***

Table 4.3: Polyserial correlations of users' self-reported *Control* over music selection and the measures of music-listening behaviour derived from their listening histories. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.4 LISTENER PROFILES

These profiles characterise music listening behaviour in terms of the user's engagement in the music retrieval. While music engagement is considered a general property of the listener, it does vary with context, and so these profiles characterise a given instance of music-listening, with a given user's engagement level perhaps varying within or between sessions.

Engaged These users have a high initial *interaction engagement* and are likely to make a selection with a high degree of *retrieval control*, with more specific retrieval queries, e.g. the selection of a particular album (Greasley, 2008). The results in this section show they then make very few further interventions, which are quick and decisive. This subsequent lack of intervention may be the result of the user having invested sufficient effort in the original retrieval, though may also be from a desire to not interfere with the album's playback, treating it as an entire work of art.

Casual These users wish to satisfice, investing little effort in the retrieval at any given point. The lack of initial *retrieval control* means that these users need to be able to easily make corrective interventions. Notably, the results in this section show that these users may ultimately make more interventions, and may actually expend more effort and have more *interaction engagement* in the retrieval over the course of an entire listening session, if measured in terms of interventions. The number of these subsequent corrective interventions could be used as relevance feedback for the recommender system, though care must be taken when interpreting behaviours in hedonic retrieval. In particular, adding such feedback loops may then influence the user's behaviour, for example the user may be concerned about how the relevance feedback will incorporate their choices and avoid the playback of embarrassing music.

Mixed Most users vary between levels of *Casual* to *Engaged* music listening. In the engagement literature in section 3.2, there are descriptions of users occupying a range of points on the continuum, and varying between these points depending upon their listening context.

5. INFORMATION-THEORETIC MEASURES

ENTROPY over music features is considered here as a metric for characterising users' music listening behaviour. This measure can be used to produce time-series analyses of user behaviour, allowing for the identification of events where this behaviour changed. In a case study, the date when a user adopted a different music retrieval system is detected. These detailed analyses of listening behaviour can support user studies or provide implicit relevance feedback to music retrieval. More broad analyses are performed across the 19,225 playlists. A Mutual Information based feature selection algorithm is employed to identify music features relevant to how users create playlists. This user-centred feature selection can sanity-check the choice of features in MIR. The information-theoretic approach introduced here is applicable to any discretisable feature set and distinct in being based solely upon actual user behaviour rather than assumed ground-truth. The techniques described here are developed to support MIR researchers in performing quantitative yet user-centred evaluations of their music features and retrieval systems.

5.1 ENTROPY OF LISTENING HISTORY

For each song being played by a user, the value of a given music feature can be taken as a random variable X. The entropy $H(X)$ of this variable indicates the uncertainty about the value of that feature over multiple songs in a listening session. This entropy measure gives a scale from a feature's value being the same for every song in the session, through to every level of the feature being equally likely across the songs. The more a user constrains their music selection by a particular feature, e.g. mood or album, then the lower the entropy is over those features. The entropy for a feature is defined as:

$$H(X) = -\sum_{x \in X} p\,(x) \log_2(p\,(x))\,, \tag{5.1}$$

where x is every possible level of the feature X and the distribution $p\,(x)$ is estimated from the songs in the listening session. The resulting entropy value is measured in bits, though can be normalised by dividing by the maximum entropy. This feature scaling achieves a variable $H'(X)$ in the range $[0, 1]$:

$$H'(X) = \frac{H(X)}{\log_2(|X|)}, \tag{5.2}$$

with the prime notation denoting the scaled function and the bar notation denoting cardinality not absolute value. The maximum entropy will in fact be determined by the minimum of either the number of possible feature levels or the cardinality of the set of songs. Estimating entropy in this way offers a generalised approach as it can be done for any set of features, though the calculation of entropy first requires that features are discretised.

Taking tempo as an example, if a user's music listening session is dominated by songs of a particular tempo, the distribution over values of a TEMPO feature would be very biased. The entropy $H(\text{TEMPO})$ would thus be very low relative to what would be expected for the overall music collection. Conversely, if users used shuffle or listened to music irrespective of tempo, then the entropy $H(\text{TEMPO})$ would tend towards the average entropy of the whole collection. It is important to consider that the music collection from which music is randomly shuffled acts as a prior distribution over the features such as TEMPO. As users control their music retrieval, the feature entropies of the retrieved music may be lower than for the overall collection, indicating the bits of information provided by the user. It is possible for the user to uniformly sample across the possible feature values (e.g. songs of a variety of TEMPO) irrespective of the collection prior, yielding a higher feature entropy than for the overall collection. The absolute value of the difference between the entropy over the collection and the entropy over the retrieved music is a measure of the information provided by the user.

APPLYING A WINDOW FUNCTION

Many research questions regarding a user's music listening behaviour concern the change in that behaviour over time. An evaluation of a music retrieval interface might hypothesise that users will be empowered to explore a more diverse range of music. Musicologists may be interested to study how listening behaviour has changed over time and which events precede such changes. It is thus of interest to extend eqn. 5.1 to define a measure of entropy which is also a function of time:

$$H(X,t) = H(w(X,t))\,, \tag{5.3}$$

where $w(X,t)$ is a window function taking W samples of X around time t. In this thesis, a rectangular window function with $W = 20$ is used, making the assumption that most albums will have fewer tracks than this. The entropy at any given point is limited to the maximum possible $H(X,t) = \log_2[n]$ i.e. where each of the W tracks has a unique value. Where W, the window size, is of fewer track plays than there are levels of the feature, the maximum possible entropy $H(X,t)$ is instead constrained by the feature levels.

An example of the change in entropy for a music feature over time is shown in Figure 5.1. In this case $H(\text{ARTIST})$ is shown as this will be 0 for album-based listening and at maximum for exploratory or radio-like listening. It is important to note that while trends in mean entropy can be identified, the entropy of music listening is itself quite a noisy signal – it is unlikely that a user will maintain a single music-listening behaviour over a large period of time. Periods of album listening (low or zero entropy) can be seen through the time-series, even after the user's adoption of radio-listening. Despite an overall trend towards shuffle or radio-like music listening, this user's history provides an example of how music listening behaviour can involve a mixture of listening styles.

CHANGE-POINTS IN MUSIC RETRIEVAL

Having produced a time-series analysis of music-listening behaviour, it is now possible to identify events which caused changes in this behaviour. In order to identify change-points in the listening history, the 'Pruned Exact Linear Time' (PELT) algorithm from Killick et al. (2012) is applied. The time-series is partitioned in a way that reduces a cost function of changes in the mean and variance of the entropy. Change-points can be of use in user studies, for example in Figure 5.1, the user explained in an interview that the detected change-point occurred when they switched to using online radio. There is a brief return to album-based listening after the change-point – users' music retrieval behaviour can be a mixture of different retrieval models. Change-point detection can also be a user-centred dependent variable in evaluating music retrieval interfaces, e.g. does the user's music listening behaviour change when they adopt a new retrieval interface?

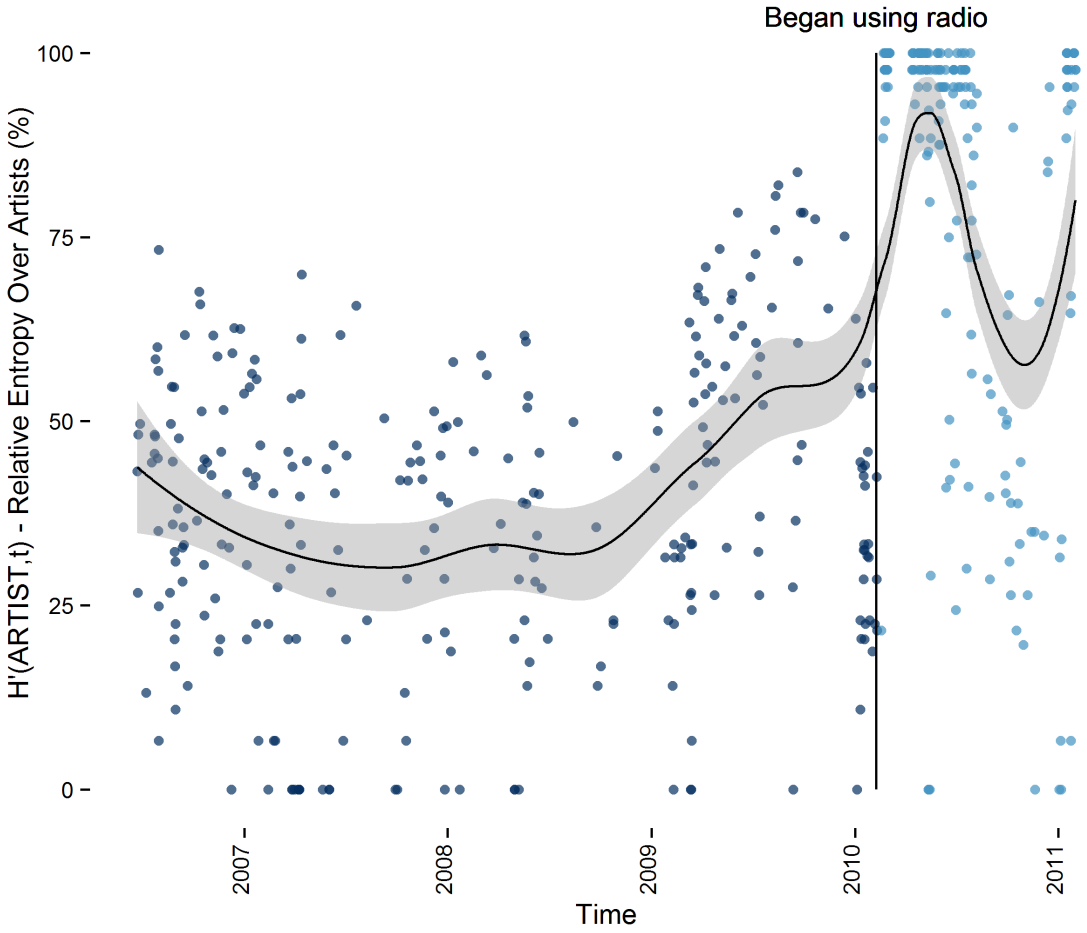

Figure 5.1: Windowed entropy view of a user's listening history. Curve fitted using locally weighted regression, shaded area represents standard confidence interval of estimated mean.

IDENTIFYING LISTENING STYLE

The style of music retrieval that the user is engaging in can be inferred using the entropy measures. Where the entropy for a given music feature is low, the user's listening behaviour can be characterised by that feature, i.e. there is a high degree of certainty about that feature's value. Alternately, where a feature has high entropy, then the user is not 'using' that feature in their retrieval. When a user opts to use shuffle-based playback, i.e. the random selection of tracks, there is the unique case that entropy across all features will tend towards the overall entropy of the music collection. In many cases, feature entropies have high covariance, e.g. songs on an album will have the same artist and similar features. Other features were not included in Figure 5.1 as the same pattern was apparent. A low entropy does not necessarily indicate that a feature is *useful* – after Ricky Martin's[1] fall in popularity, most music listening sessions probably have a low entropy for the LATIN music feature of genre classifiers.

[1] A singer popular in the 1990s who gave exposure to the Latin music genre.

5.2 USER-CENTRED FEATURE SELECTION

Identifying which music features best describe a range of playlists is not only useful for playlist recommendation, but also provides an insight into how users organise and think about music. This section develops an approach to ranking music features according to the information they share with playlist organisation. Commercial features are used as an example, including high level features such as genre, mood and popularity. As the existing music retrieval systems used by listeners are based upon these features, it is likely that a 'chicken-and-egg' effect will apply, where the features which best describe user playlists are those which users are currently exposed to in their existing retrieval interfaces. This approach requires discretised features, and is sensitive to the choice of discretisation strategy. The examples shown use features that were already discretised by the commercial provider.

MUTUAL INFORMATION

Information-theoretic measures can be used to identify to what degree a given feature shares information with class labels. For a feature X and a class label Y, the mutual information $I(X;Y)$ between these two can be given as:

$$I(X;Y) = H(X) - H(X\,|\,Y)\,, \tag{5.4}$$

that is, the entropy of the feature $H(X)$ minus the entropy of that feature if the class is known $H(X\,|\,Y)$. Taking membership of playlists as a class label, it is then possible to determine how much can be known about a song's features if one knows what playlist the song is in. When using mutual information to consider class labels in this way, care must be taken to account for random chance mutual information (Vinh et al., 2010). This approach is adapted to focus on how much the feature entropy is reduced, and is normalised against the expectation of this feature's entropy accordingly:

$$AMI(X;Y) = \frac{I(X;Y) - E[I(X;Y)]}{H(X) - E[I(X;Y)]}, \tag{5.5}$$

where $AMI(X;Y)$ is the adjusted mutual information and $E[I(X;Y)]$ is the expectation of the mutual information, i.e. due to random chance. The AMI gives a normalised measure of how much of the feature's entropy is explained by the playlist. When $AMI = 1$, the value of the given feature for each track is known exactly if the playlist is known. When $AMI = 0$, nothing about the feature is known if the playlist is known. Those features that have a high AMI with playlist membership are related to the way in which user's organise their music, and are thus of greater interest when designing and evaluating music retrieval systems.

LINKING FEATURES TO PLAYLISTS

The 19,225 playlists in the *SPUD* dataset were used to calculate their AMI with a variety of high level music features from Syntonetic and Spotify. The ranking of some of these features is given in Figure 5.2. The aim here is merely to illustrate this approach, as any results are only as reliable as the underlying features. With this in mind, the features ROCK and ANGRY had the most uncertainty explained by playlist membership. While the values may seem small, they are calculated over many playlists, which may combine moods, genres and other criteria. As the features with a high AMI change most between playlists (rather than within them), they are the most useful for characterising the differences between the playlists. The DURATION feature ranked higher than expected, further investigation revealed playlists that combined lengthy DJ mixes. It is perhaps unsurprising that playlists were not well characterised by whether they included features such as WORLD music, which is a 'miscellaneous' category. It is of interest that TEMPO was not one of the highest ranked features, illustrating the style of insights available when using this approach. Further investigation is required to determine whether playlists are not based on tempo as much as is often assumed or if this result is due to the peculiarities of the proprietary perceptual tempo detection.

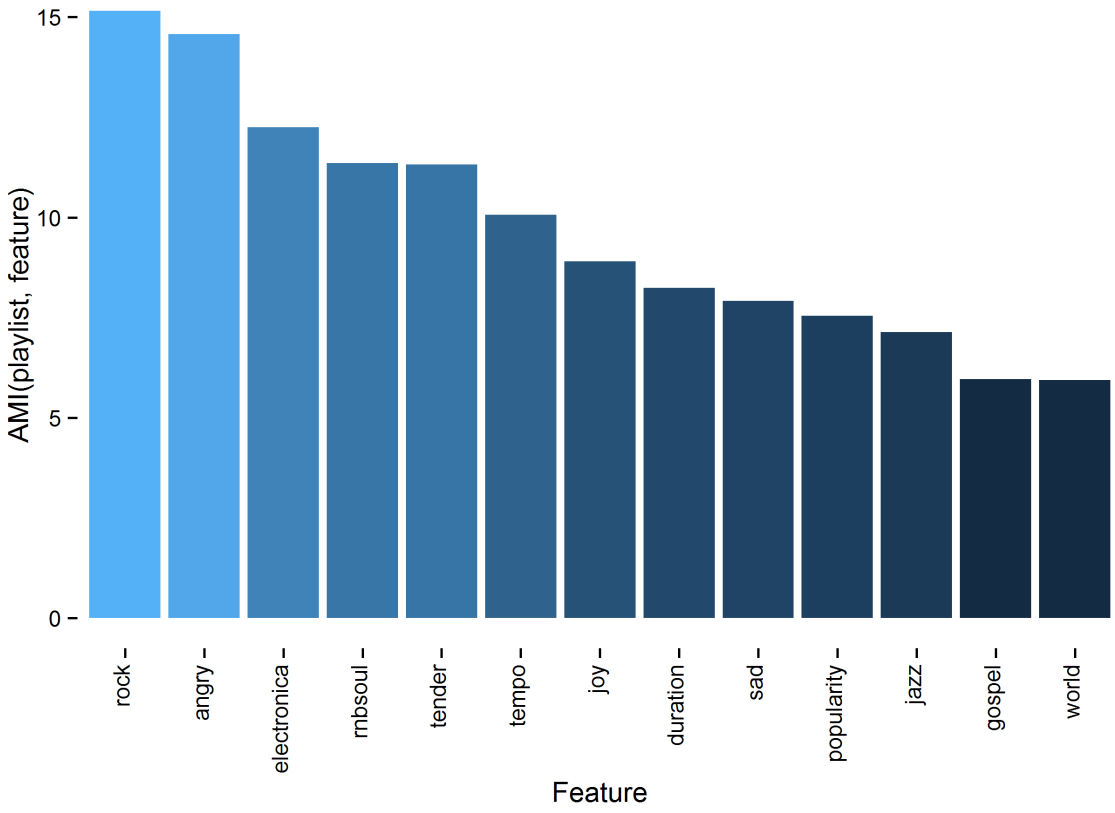

Figure 5.2: Features are ranked by their Adjusted Mutual Information with playlist membership. Playlists are distinguished more by whether they contain ROCK or ANGRY music.

FEATURE SELECTION

Features can be selected using information-theoretic measures, with a thorough treatment of this field given by Brown et al. (2012). They introduce a unifying framework within which to discuss methods for selecting a subset of features, using their J criterion for a feature f_n and (playlist) class C given the previously selected feature subset S:

$$J\left(f_n\right) = I(f_n; C \mid S)\,. \tag{5.6}$$

This gives a measure of how much information the feature shares with playlists given some previously selected features, and can be used as a greedy feature selection algorithm. Intuitively, features should be selected that are relevant to the classes but that are also not redundant with regard to previously selected features. A range of estimators for $I(f_n; C \mid S)$ are discussed in Brown et al. (2012).

As a demonstration of the feature selection approach described, it is applied to the features depicted in Figure 5.2, selecting features to minimise redundancy. The selected subset of features in rank order is: ROCK, DURATION, POPULARITY, TENDER and JOY. It is notable that ANGRY had an AMI that was almost the same as ROCK, but it is redundant if ROCK is included. Unsurprisingly, the second feature selected is from a different source from the first – the duration information from Spotify adds to that used to produce the Syntonetic mood and genre features. Reducing redundancy in the selected features in this way yields a very different ordering, though one that may give a clearer insight into the factors behind users' construction of playlists.

5.3 DISCUSSION

The feature selection shown in this chapter is done directly from the user data. In contrast, feature selection is usually performed using classifier wrappers with ground truth class labels such as genre. The use of genre is based on the assumption that it would support the way users currently organise music and features are selected based on these labels. This has led to issues including classifiers being trained on factors that are confounded with these labels and that are not of relevance to genre or users (Sturm, 2014a). The user-centred approach introduced here selects features independently of the choice of classifier, in what is termed a 'filter' approach. The benefit of doing this is that a wide range of features can be quickly filtered at relatively little computational expense. While the classifier 'wrapper' approach may achieve greater classifier performance, it is more computationally expensive and more likely to suffer from overfitting (as common in genre classification for example).

The key benefit of filtering features based on user behaviour is that it provides a perspective on music features that is free from assumptions about users and music ground truth. This user-centred perspective provides a sanity-check for music features and classification – if a feature does not reflect the ways in which users organise their music, then how useful is it for music retrieval?

WHEN TO LEARN

The information-theoretic measures presented offer an implicit relevance feedback for users' music retrieval. While this chapter has considered the entropy of features as reflecting user behaviour, this behaviour is conditioned upon the existing music retrieval interfaces being used. For example, after issuing a query and receiving results, the user selects relevant songs from those results. If the entropy of a feature for the songs selected by the user is small relative to the full result set, then this feature is implicitly relevant to the retrieval.

The identification of shuffle and explorative behaviour provides some additional context for this implicit relevance feedback. Music which is listened to in a seemingly random fashion may represent an absent or disengaged user, adding noise to attempts to weight recommender systems or build a user profile. At the very least, where entropy is high across all features, then those features do not reflect the mental model currently being employed by the user for their music retrieval. The detection of shuffle or high-entropy listening states thus provides a useful data hygiene measure when interpreting listening data.

ENGAGEMENT

The entropy measures capture how much each feature is being 'controlled' by the user when selecting their music. It has been shown that it spans a scale from a user choosing to listen to something specific to the user yielding control to radio or shuffle. Considering entropy over many features in this way gives a high-dimensional vector representing the user's engagement with music. Different styles of music retrieval occupy different points in this space, commonly the two extremes of listening to a specific album or just shuffling. There is an opportunity for music retrieval that has the flexibility to support users engaging and applying control over music features only insofar as they desire to. An example of this would be a shuffle mode that allowed users to bias it to varying degrees, or to some extent, the feedback mechanism in recommender systems. This information-theoretic characterisation of listening behaviour and user control provides further quantitative grounding to the discussion of engagement and control in section 3.2.

5.4 CONCLUSIONS

This chapter has proposed the use of information-theoretic measures to capture users' music-listening behaviour and relate it to an arbitrary feature set. Some example applications were presented, including an analysis of a user's listening history in terms of changes in the music features, and identifying the degree to which music features shared information with the way users organised playlists. These approaches are highly generalisable, any clustering could be investigated instead of playlist membership, for example the time of day or location of track plays. An interactive example of Figure 5.1, where parameters such as window size can be changed, is available along with the corresponding code.[2]

The measures proposed in this chapter must still be interpreted in the context of a user-centred evaluation. Change-points in the entropy of a user's music listening can identify important events to lead discussions with users, however would be difficult to interpret by themselves. Using relative entropy can go some way towards accounting for the differing nature of users' music collections. In the playlist example, there is a question as to what the Adjusted Mutual Information should be calculated in relation to. Showing AMI relative to to the feature entropy is an intuitive way to present the amount of shared information. It is usual and typically more appropriate, however, to calculate AMI relative to the maximum of either the class or feature entropy. The distinction is whether a feature capturing relatively few bits of information should be ranked highly if all of that information is explained by playlist membership.

While the features are described as being from Syntonetic and Spotify, further information regarding the features is limited due to commercial sensitivity. The choice of features is such that they are high-level and intuitive, serving as placeholder examples which can be substituted with other music features. The use of commercially-derived features is because of licensing issues with large music collections.

Measuring the information content of a retrieval interaction provides a quantitative grounding to the idea of *retrieval control*. As discussed in chapter 3, the reduction of entropy from an overall music collection in making a selection gives a measure of how controlled the selection was. Making the distinction about whether the bits of information were provided by the user or an autonomous recommender system can represent the handover of control. These ideas are applied in the evaluation of the commercial BeoSound Moment product in chapter 8, capturing the entropy of recommendations designed to match user engagement.

[2] http://www.dannyboland.com/spud/

Part III

Designing for Engagement in Music Retrieval

6. QUERY BY TAPPING

RHYTHM is a musical universal. This chapter explores the use of rhythmic queries, with a view towards creating a low–engagement music retrieval interaction. As aspects of rhythmic perception are universal across listeners (as noted in section 2.6), this chapter explores the hypothesis that it should be possible to create a *Query by Tapping* (*QbT*) music retrieval system with a high degree of correspondence with users' mental models of query production. A study of rhythmic querying is conducted, finding a surprisingly high degree of subjectivity in how users reproduce musical rhythm, with subsequent development of a generative model of rhythmic queries. This model is then incorporated into a demo music retrieval system, along with novel methods for interpreting rhythmic queries and inferring relevant musical works. The system is implemented on a mobile device, to allow users to search for a specific song by tapping its rhythm or general tempo onto the device, detected by capacitive touchscreen or accelerometer. The generative query model can be trained to an individual user, capturing their style of rhythmic query, and underpins a user-centered approach to *QbT* systems. An experiment is conducted, to demonstrate the benefit of the use of a trained generative model over existing approaches. Feedback from participants is then acquired, being generally positive and including a suggested use case for in-pocket mobile interaction.

The way listeners comprehend and reproduce rhythm is fundamental to music and universal across cultures, as discussed in section 2.6. Tapping along to the rhythm of a song is a common sign of listener engagement with music. Exploiting this predisposition to tapping rhythm as a form of music retrieval would allow for music selection on a device with very limited sensing ability – a microphone, button or single capacitive sensor would suffice. As a large proportion of music listening occurs in mobile contexts, and mobile devices are now instrumented with a variety of sensing capabilities, mobile listening is a compelling use case for *QbT*. In particular, mobile *QbT* would free users from having to remove their mobile music player from their pocket when influencing their music playback.

In Saponas et al. (2011), capacitive sensors were able to detect touch input through fabric, supporting gestural input including drawn letters. While this input modality could support explicitly 'typing' a music query, this would require a high degree of *interaction engagement* from users. Users would have to engage in a more tightly-coupled interaction loop to type a query than would be needed with rhythmic querying, i.e. having to think of an exact track and spelling it rather than casually tapping a beat. Manabe and Fukumoto (2012) showed that tapping input can now also be detected via headphones, making it an ideal input modality for mobile music-listening contexts. Minimising the technological footprint of an interaction in this way not only lowers cost but also frees designers from the encumberance of integrating displays, keyboards etc. into a music retrieval system.

Existing Efforts

The retrieval of a musical work by tapping its rhythm is a problem that has received some consideration in the Music Information Retrieval community. The term was introduced in Jang et al. (2001), which demonstrated that rhythm alone can be used to retrieve musical works, with their system yielding a top 10 ranking for the desired result 51% of the time. Their music corpus consisted of MIDI representations of tunes such as *You are my sunshine*, which is unlikely to represent the type of popular music a user would wish to retrieve. The work was also limited in considering only monophonic rhythms, i.e. the rhythm from only one instrument, as opposed to being polyphonic and comprising of multiple instruments. The failure to consider polyphonic rhythm (comprising of multiple voices/instruments) is a major limitation of prior *QbT* systems. Such systems use one rhythmic sequence as being the de facto rhythm for a musical work, requiring that all users tap a musical rhythm in the same way. This issue extends even to evaluation, with queries used for *QbT* evaluations in MIREX generated by people trained to produce rhythm a particular way. After the publication of the work in this chapter (Boland and Murray-Smith, 2013), Kaneshiro et al. (2013) built a *QbT* dataset from trained participants that is now used in MIREX, and noted that participants "wanted to tap their instrument's part instead."

The potential of rhythmic interaction has been recognised in Human Computer Interaction (Lantz and Murray-Smith, 2004; Wobbrock, 2009), for example with Ghomi et al. (2012) introducing rhythmic queries as a replacement for hot-keys. In Crossan and Murray-Smith (2006), tempo is used as a rhythmic input for exploring a music collection – indicating that users enjoyed such a method of interaction. The consideration of human factors is also an emerging trend in Music Information Retrieval, as discussed in chapter 2. This work draws upon both these themes, being the first *QbT* system to adapt to users.

A number of key techniques for *QbT* are introduced in Hanna and Robine (2009), which describes rhythm as a sequence of time intervals between notes – termed *inter-onset intervals (IOIs)*. They identify the need for such intervals to be defined relative to each other to avoid the user having to exactly recreate the music's tempo. In previous implementations of *QbT*, each *IOI* is defined relative to the preceding one (Hanna and Robine, 2009). This sequential dependency compounds user errors in reproducing a rhythm, as an erroneous *IOI* value will also distort the subsequent one. The approach to rhythmic interaction in Ghomi et al. (2012) however used k-means clustering to classify taps and *IOIs* into three classes based on duration. Their clustering based approach avoids the sequential error however loses a great deal of detail in the rhythmic query. Applying a clustering approach to musical queries would need a greater number of clusters to be identified.

ONSET DETECTION

In order to compare user queries against a music library, one must compute the intervals (*IOIs*) between the rhythmic events within the music. Onset detection is the task of finding such events and has been studied at length within the field of Music Information Retrieval. An evaluation of onset detection algorithms in Böck et al. (2012) showed the most precise onset detection method reviewed was their variant of the 'spectral flux' technique introduced by Masri (1996), which measures how quickly the power spectrum of a signal is changing. They also discuss the benefits of adaptive whitening introduced in Stowell and Plumbley (2007) which adaptively normalises frequency bands' magnitudes to improve onset detection in music with highly varying dynamics, such as the rock music used in this work.

These off-the-shelf onset detection techniques are applied to the audio of the tracks users could query, and update the existing work on *QbT* to a state-of-the-art implementation. The onsets acquired from audio are used as a baseline, along with professionally annoted music onsets. The *QbT* techniques developed in this chapter are applied to both the audio-derived and annotated onsets, and the evaluations are conducted using both. Though the comparison with the audio-derived onsets provides a context within which to view the potential results shown using annotated onsets, it is the comparison between the results using the annotated onsets that provide an insight into the developments in this chapter.

Figure 6.1: Participants entered rhythmic queries by tapping the capacitive touchscreen of a Nokia N9 mobile phone. The timing of these taps was logged, to enable comparison between users' queries for a given song, as well as against annotated musical scores.

6.1 INITIAL STUDY

An exploratory study was conducted, eliciting rhythmic queries from participants to explore the feasibility of music filtering using rhythmic queries. 10 European students were recruited as participants to produce rhythmic queries of songs, which they selected from a corpus of 1000 songs. The corpus was collected from participants' own MP3 music collections and combined, with the same corpus then presented to all participants. These music files were also used to obtain the *IOIs* from audio, using the state-of-the-art onset detection techniques discussed previously. The rhythmic queries were entered by the participant tapping on the touchscreen of a Nokia N9 (as in Figure 6.1), which had been configured to record the time intervals between taps. The queries provided by the users were compared using the techniques described later in section 6.2. For this initial study, the phone screen remained blank. Users were instructed to select a song from a printout of the files in the corpus and then to "tap the rhythm of the song on the touchscreen, in order to select that piece of music."

One aim of this initial data collection was to investigate the way in which the participants produced their rhythmic queries, without guidance from the experimenter. This is in contrast to previous approaches, where rhythmic query onsets were acquired after instructing users to only reproduce the vocal lead track. Though informal, this query collection study enabled a discussion with users about how they were producing their queries, as well as follow-up questions post-hoc.

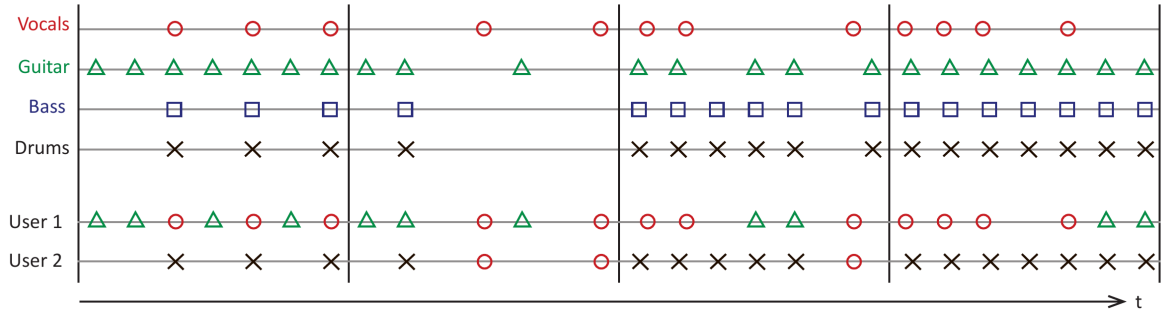

Figure 6.2: An illustration of how users construct queries by sampling from preferred instruments. User 1 prefers Vocals and Guitar whereas User 2 prefers Drums and Vocals.

INTER-PARTICIPANT VARIABILITY

As an initial sanity check, the queries produced by multiple participants for the same song were compared against each other. Surprisingly, little similarity was identified for a large number of the songs. In discussions with the participants, it became apparent that a variety of strategies were employed when annotating the rhythm of a piece of music. In particular, participants identified particular instruments which they would entrain with – annotating those instruments' onset events when available. Participants were also not equally detailed when producing the queries, with some using fewer taps than others to represent the same rhythm in a piece of music. A depiction of how participants might sample from the available instrument onsets is given in figure 6.2. A well known but nonetheless significant further complication is that users often query from different parts of a song, and this behaviour was also apparent amongst the participants.

The observations made in this initial study indicate that music retrieval using rhythmic queries is much more complex than was originally expected. In particular, there is a need for learning the user-specific habits in producing rhythmic queries. The conversations with participants provided some insight into how their tapping strategy may be modelled, with their affinities for the various instruments and their verbosity identified as important features. Instrument affinity can be captured as a list of the available instruments, ordered in terms of priority. Participants stated that they switched instruments when their preferred instrument became available. Query verbosity is more difficult to model, the next section turns again to the music psychology literature to identify 'referent period' as a means of capturing the user's degree of verbosity.

The work in this chapter explores the use of these features to construct a model of how users generate queries. Such a *generative* model can address the variance in user query production and allow input queries to be predicted and thus matched. An alternative approach however would be to provide instruction or feedback to the user about how to tap rhythmic queries in a way that the system understands.

6.2 EXEMPLAR SYSTEM

It is common when interacting with music systems to be presented with a list of retrieved music, perhaps as a playlist, which is then played sequentially. Rhythmic queries can be used to infer a belief over a music collection about which songs the user wishes to listen to. This section documents the development of a mobile interaction that allows a user to 'shake up' a list of music by tapping the desired rhythm onto their device, with the music then being sorted by rhythmic similarity. The user can then play through the resulting playlist arranged by these features or proceed to select their intended song. One example use case is shown in Figure 6.3, where a user enters a rhythmic query without removing the mobile device from their pocket. This interaction highlights the flexibility of rhythmic queries, allowing users to find songs of a given tempo or with certain rhythmic properties or to simply select a specific song. A mobile demonstrator system is implemented and evaluated quantitatively as well as qualitatively with Singaporean users, demonstrating its viability and cross-cultural application.

Figure 6.3: Music can be shuffled by tapping a rhythm on the mobile device, allowing for casual music interaction.

EXPOSING SYSTEM BELIEF & UNCERTAINTY

Displaying a ranked list of songs would lose some of the information about the user's songs of interest, which the system has inferred from matching the user's query to the music collection. For example several songs may have a very similar level of belief held about them and this would not be communicated by simply displaying a sorted list. By exposing the uncertainty in the interaction to the user through some visual feedback, they will be better able to understand the behaviour of the system and and develop an appropriate mental model for producing the most discriminative queries.

To better expose the beliefs held by the system, the size of each list entry is scaled by this belief. If one song alone is a particularly strong candidate then its font size will be much larger than the other entries. Similarly, where there is uncertainty across a number of songs, these will be a similar size – making the user aware of the uncertainty within the interaction. This scaling of the tracks of interest not only reflects the retrieval result but allows for the easier selection of the intended track. This feedback also reflects the uncertainty in the retrieval, from both the user's query as well as similarity in the music collection. Such feedback of uncertainty can be applied generally, for example distorting a music map based on a belief over the music space or scaling the words in an word cloud of musical artists or genres.

Figure 6.4: Music playlist initially sorted alphabetically (left) and after a query for an upbeat rock song "Any Way You Want it" (right). The font size of the intended track is scaled larger as a belief of the user's interest in it is inferred, and unlikely songs are scaled smaller. The order of the tracks is also changed according to the inferred belief, allowing the retrieval to act as a constrained shuffle.

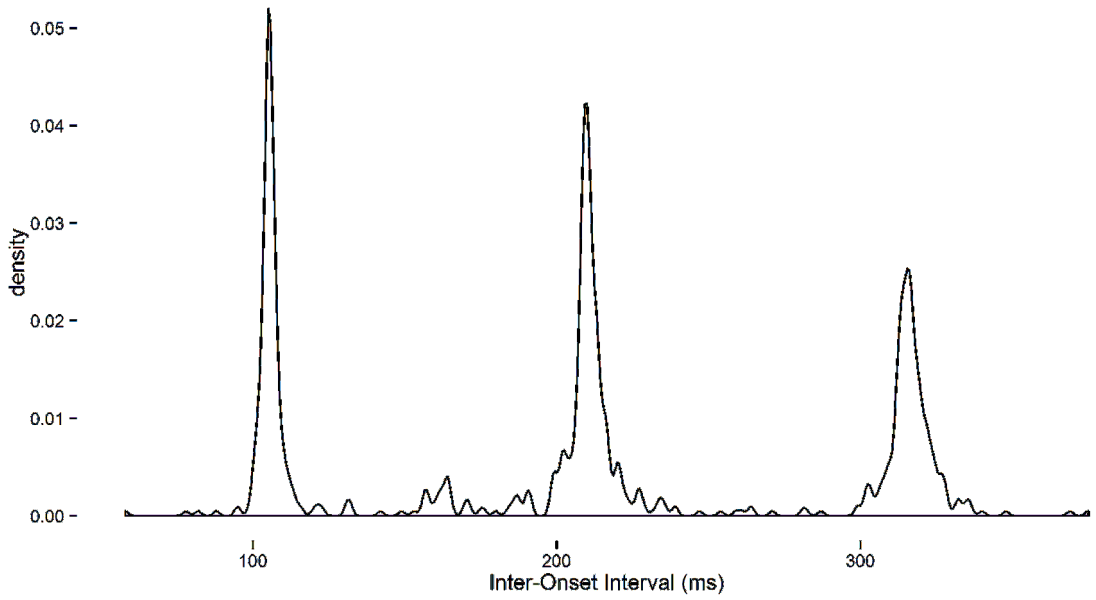

Figure 6.5: Kernel density plot of IOIs in a rhythmic query, showing the clustering around categories of IOI values. Note that the mean of each category is a multiple of the smallest, e.g. 110, 220, 330mS giving a tatum of 110mS.

INTERPRETING RHYTHMIC QUERIES

As discussed in section 2.6, listeners perceive and reproduce rhythm in relative categories of intervals between notes (*IOIs*). These intervals are constructed as multiples of some reference, greatest common divisor beat termed the *tatum* (Seppänen, 2001). Complex rhythms are thus distorted into distinct categories of *IOIs*, each defined relative to each other. The tatum for the query must be established in order to then interpret the rhythmic pattern (as distinct from the absolute position of taps in the time domain). The tatum is estimated in this work by taking the autocorrelation of the histogram of *IOI* categories, giving the reference unit in which they can be defined. One must take this approach of defining rhythmic events relative to each other as users cannot accurately reproduce the absolute timing of music. Taking only the relative timings would discard some information however, such as the tempo the query was produced at. To avoid this, the tatum itself can be used as an additional feature for weighting the retrieval.

For controlled non-musical rhythms, Ghomi et al. (2012) used k-means clustering to identify long and short interval clusters. As it is expected that users will generate *IOIs* by sampling from distributions around an unknown number of *IOI* categories, an alternative approach of modelling the categories is to fit a Gaussian Mixture Model selected using the Bayes Information Criterion. An example of the clustering is described in figure 6.5.

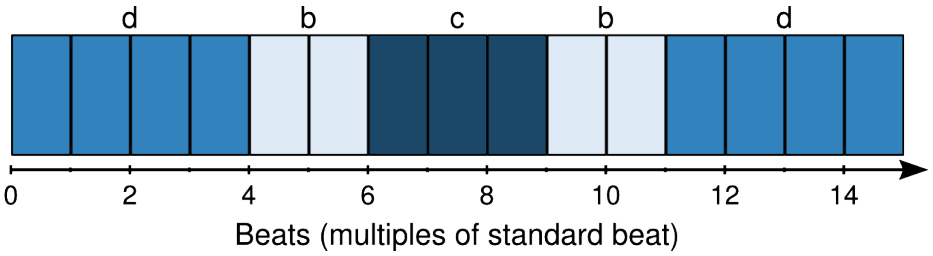

Figure 6.6: A rhythmic query depicted with alternately coloured intervals. The mapping between interval duration and string characters is shown.

RHYTHMIC STRING MATCHING

The *IOI* categories are assigned labels 'A', 'B' etc. for convenience. This allows the rhythmic queries to be easily encoded as a string of *IOI* category labels. For example, an interval double the length of the tatum would be classified as 'B'. An example query is depicted in figure 6.6, showing the mapping from interval to string character.

The problem of matching rhythm can now be generalised to string matching, for which many efficient algorithms exist. As in Hanna and Robine (2009), the Smith-Waterman local alignment algorithm is used, as this is able to match a query against any part of a song. Similarly, the algorithm was adapted to scale the penalty with the mismatch error. An advantage of this approach over that in Hanna and Robine (2009) is that no thresholds are required, with the mismatch error being proportional to the difference between normalised *IOIs*:

$$E_{\text{IOI}} \propto \left(\text{IOI}_{\text{query}} - \text{IOI}_{\text{song}}\right).$$

A further feature of the Smith and Waterman (1981) algorithm is that it was developed to allow for gaps in sequences, in order to match spliced DNA sequences. This feature is also useful in *QbT*, as sections of a song in which the generative model fails to correspond to the user's query will simply be considered as a gap and given a limited penalty. The cost function of the string matching algorithm can assign a different score S for matched, missing or incorrect *IOIs*. A parameter G weights the penalty for a gap, such that a gap is equivalent to G *IOI* mismatches. In this work $G = 2$, assuming that if a query has two consecutive mismatched *IOIs* then the query is no longer conforming to the model at that point. The penalty scores are calculated as follows:

$$\begin{aligned} S_{\text{match}} &= 10, \\ S_{\text{mismatch}} &= -abs(IOI_A - IOI_B), \\ S_{\text{gap}} &= -G \times S_{\text{match}}. \end{aligned}$$

$$H = \begin{pmatrix} & C & C & B & D & C & B & C \\ B & 0 & 0 & 10 & 0 & 0 & 10 & 0 \\ C & 10 & 10 & 0 & 0 & 10 & 0 & 20 \\ C & 10 & 20 & 0 & 0 & 10 & 0 & 10 \\ B & 0 & 0 & 30 & 10 & 0 & 20 & 0 \\ D & 0 & 0 & 10 & 40 & 20 & 0 & 10 \\ A & 0 & 0 & 0 & 20 & 20 & 10 & 0 \\ C & 10 & 10 & 0 & 0 & 30 & 10 & 20 \\ B & 0 & 0 & 20 & 0 & 10 & 40 & 20 \\ C & 10 & 10 & 0 & 10 & 10 & 20 & 50 \end{pmatrix}$$

Figure 6.7: The Smith-Waterman algorithm compares a query against a target sequence, matching 'CCBD-CBC'.

The algorithm constructs an $n_{\text{query}} \times m_{\text{song}}$ matrix H (as in figure 6.7) where n is query length and m is the target sequence length. If the strings were identical then the diagonal of the matrix would identify each matching character pair, thus diagonal movements incur no penalty. In the example shown, one sequence has an 'A' removed (the downward step) to give a better match and thus a penalty is deducted from the score. Penalties are assigned when the other movements are required in order to create a match, with a back-tracking process used at the end to find the (sub)path with the least penalty. This process allows for the best matching subsequences to be identified – in this work, a query matched against a larger song.

TATUM AS A FEATURE

Previous work on *QbT* defines the rhythm irrespective of tempo (or tatum), as is done here. It has been shown however that tempo can be a useful feature in browsing a music collection (Crossan and Murray-Smith, 2006). Tatum (being related to tempo) should be used as an additional feature to weight the ranking of rhythmic queries. The weighting given to this feature could additionally be adapted to each user, though that is not explored in this work. The tatum t error function is defined logarithmically, so that halving a duration is equivalent to doubling it:

$$E_{\text{t}} = \left(\log_2 \left(\frac{\text{t}_{\text{query}}}{\text{t}_{\text{song}}} \right) \right)^2 .$$

The tatum error is used as a prior over the music space when performing the rhythmic string comparison, biasing the results to those with similar tatum values. This helps discern amongst songs which are temporally very different but which share a similar rhythmic pattern. Where users only wish to listen to a particular style of music or cannot recall the rhythm of a song, they can simply tap a query at a desired tempo. If the rhythmic events are equally spaced (as in a metronome) then only the tatum is used to discriminate amongst the songs.

It is worthwhile to note that the tatum is not necessarily the inverse of tempo. Tempo is often calculated as 'beats per minute', with an average value acquired across all the rhythmic events. It follows from this that the measured tempo would be highly dependent upon the section of song used to produce a query. The tatum however is the base unit that all the rhythmic events are multiples of and should be more stable throughout a piece of music. This distinction is largely only of interest from a technical perspective and generally, the tatum is inversely proportional to tempo.

6.3 GENERATIVE MODEL

Rhythmic queries for a given song can vary greatly between subjects though are typically consistent within subjects. In order to build a database against which rhythmic queries can be matched, a generative model is required that can account for this variability. The use of the generative model encodes the knowledge about user behaviour obtained from the initial study. In essence, the model is designed to answer the question "*What would the user do?*" to achieve an outcome (selecting a target song). Training the model to users can be done by setting a fixed outcome and asking users to provide the input they would provide to achieve that outcome. The inference of the user's intended music is conditioned entirely upon the model and so should inherently improve as the generative model is improved or trained.

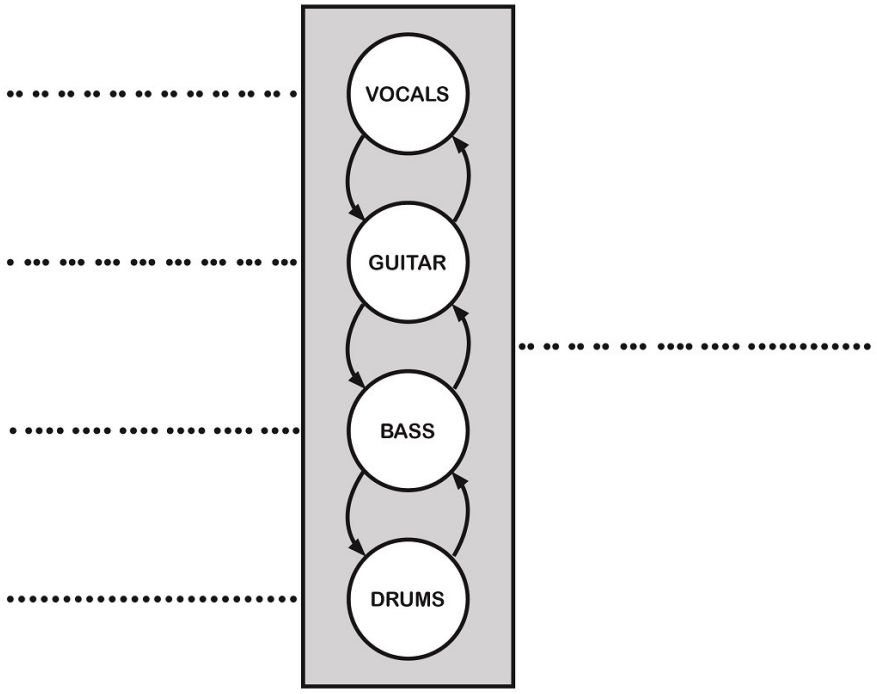

Figure 6.8: The generative model samples onset events from multiple instrument streams, producing a single output sequence.

INSTRUMENT AFFINITY

Polyphonic music will have a sequence of notes for each instrument and thus a sequence of *IOIs* for each instrument. As observed in the initial study, users typically switch between instruments as they feature in the music. This switching behaviour follows the user's preference of instruments to tap to. This set of preferences for instruments is termed here as the affinity vector A_{ff}, which ranks the available instruments in terms of the user's affinity. The generative model uses A_{ff} to switch to a preferred instrument's *IOI* sequence as those instruments feature. The behaviour of this model can be considered as a finite state transducer with a state for each instrument, sampling from the instrument sequence corresponding to the current state, as in figure 6.8.

Users are able to anticipate upcoming musical notes and will not perform the switch to another instrument state if their preferred instrument sequence will shortly resume. This look-ahead behaviour is also implemented in the generative model. The model samples notes up to $500mS$ in advance and stores them in a look-ahead buffer; only when the buffer is empty does the state change to the next available in the affinity vector. Whenever a note event occurs for an instrument with a greater affinity, the model immediately changes state. For ease of calculation, the model is treated deterministically and a database of reference rhythmic sequences is produced across the music collection. A probabilistic approach to switching may better model user behaviour but would add a great deal of complexity.

REFERENT PERIOD

As music is highly structured, rhythm can be thought of as a hierarchy, where a note on one level could be split into multiple notes on a lower level. Individuals have a 'referent period', i.e. a rate at which information processing is natural to them and they are likely to synchronize at a level in the hierarchy of musical rhythm that is closest to their referent period. It has been shown that musical training and acculturation result in higher referent levels (Drake and El Heni, 2003), and so it is likely that rhythmic queries will be produced from a variety of levels in the rhythmic hierarchy, depending on the issuing user. This adds an extra degree of complexity to the generative model, in that the generated queries must match the referent period with which the user produces queries.

In order to model the differing referent periods of users, a music corpus must contain onset data for several levels of rhythmic complexity. The appropriate level can then be selected as part of training the generative model. Such training enables the acquisition of a prior over possible referent periods for the user – it may be the case that the user's referent period varies between songs.

MODEL IMPLEMENTATION

A model was implemented to recreate users' onsets sampling strategy, as discovered in the initial study and depicted in Figure 6.2. The implementation of this sampling strategy is detailed as pseudocode in algorithm 6.1 and also shown graphically in Figure 6.8. While this model is able to generate canonical query sequences for a given instrument affinity A_{ff}, the queries produced by users will likely be much more varied. The string matching described in subsection 6.2 is used to calculate the distance of a query from the canonical queries produced by this model. These distance values are then used to approximate the probability of a given query being produced for a set of model parameters (forming the likelihood function in section 6.3). The user model parameters are also estimated using the string matching distances, section 6.3 describes the use of of training queries matched to the canonical queries to infer a distribution over the possible parameters.

Data: Voices - An array containing streams of note onset times for each voice in a piece of polyphonic music. A_{ff} - A sorted array of voice indices, indicating the user's ranked affinity for entraining with each voice.

Result: A stream of note onsets for this song, as would be tapped by a user.

```
t=0;
state=None;
Order Voices by A_ff indices;
while more note onsets to sample in Voices do
    for v ∈ Voices do
        if note available in v within 200ms of t then
            sample next note;
            set state to v;
            break;
        else if state = v and note available in v within 500ms of t then
            /* Note soon, so don't check other voices          */
            break;
    end
    if note sampled then
        output note;
    increment t;
end
```

Algorithm 6.1: Sampling from multiple polyphonic voices according to Affinity vector (user's preference for entraining with different instruments).

INFERRING USER INTENT

The task of ranking songs based on some rhythmic evidence can be seen as an inference task and not only as a traditional retrieval task. Previous work in information retrieval has introduced the use of query models to encode knowledge about how a user produces a query (Lafferty and Zhai, 2001). The approach here is similar in the use of a query likelihood model. When producing a rhythmic query, the user uses their internal query model $\vec{M}_u$. They then produce a query using this model, which is matched against the music corpus. The problem can thus be expressed using Bayes' theorem:

$$p\left(d_j \,|\, q, \vec{M}_u\right) = \frac{p\left(q \,|\, d_j, \vec{M}_u\right) p\left(d_j \,|\, \vec{M}_u\right)}{p\left(q \,|\, \vec{M}_u\right)}.$$

That is, one can infer a belief about the intended song conditioned upon the query q by computing the likelihood of the query being produced for each song d_j in the music space. The prior $p\left(d_j \,|\, \vec{M}_u\right)$ should be non-informative, currently there is no evidence that music listening intent is directly conditioned upon the user's query model for tapping to music. In order to perform the above inference, the generative model of user queries $\vec{M}_u$ must be trained to the user. This is done by taking a fixed outcome (i.e. selecting a target song d_t) and asking users to provide a suitable query q so as to achieve that outcome. The parameters of the generative model are then inferred from the query:

$$p\left(\vec{M}_u \,|\, q, d_t\right) = \frac{p\left(q \,|\, d_t, \vec{M}_u\right) p\left(\vec{M}_u \,|\, d_t\right)}{p\left(q \,|\, d_t\right)}.$$

In this work the prior $p\left(\vec{M}_u \,|\, d_t\right)$ is non-informative however it is probable that the user's approach to tapping music is conditioned upon the particular song to some extent. In wider use where a large corpus of queries has been collected, it would be possible to compute a prior belief about the tapping model used for each song. This should improve the inference of the user's general tapping model. For the work here the model is trained for a given song and a given participant to account for this, as well as looking at training across songs for a subset of participants.

In order to infer a belief about whether a user is interested in a given song, one must compute the likelihood $p\left(q \,|\, d_j, \vec{M}_u\right)$ of their query conditioned upon that song being of interest and that user's query model. The string matching function is used to compare user queries with those in the database and to assign beliefs to songs accordingly. The more edits that are required to match the query to the stored song sequence, the lower the estimated likelihood of that query for that song.

6.4 EVALUATION

A within-subjects experiment was conducted to compare the performance of the query likelihood model described before and after training, as well as against a baseline approach using the onset detection techniques. The baseline approach used the onset detection techniques described in section 6 and reviewed by Böck et al. (2012). These conditions are termed: *Baseline* (onset detection), *Untrained GM* (polyphonic data with generative model) and *Trained GM* (polyphonic data with trained generative model).

Given the parameters of the generative model, the target space against which queries are matched is greater than in the baseline case. For the baseline, there is one possible sequence for each of the 300 songs. The model has four instrument sequences for each song, sampled using the generative model with 96 possible parameter permutations, yielding a target song space of $28,800$ sequences.

EXPERIMENTAL SETUP

A corpus of MIDI and MP3 music data was acquired from popular rhythm games, featuring professionally annotated note onset times (from which *IOIs* are computed) for each instrument in 300 rock and pop songs. While the size of this corpus was limited by the source of the data, it reflected real-world usage at the time – Karaganis and Renkema (2013) gives it as the median music file collection size in Germany. Participants selected at least two songs from the corpus and listened to them to ensure familiarity. They were then asked to produce at least three rhythmic queries for each song by tapping a section of the song's rhythm on the touchscreen of a Nokia N9 mobile phone. No feedback was provided to the participant after each query. The queries were used to train the generative model using leave-one-out cross-validation. Participants were provided with headphones to control background noise.

Quantitative data was captured in the form of rank results, with songs ranked according to the inferred belief. Qualitative data was captured during a discussion with participants where they were asked to comment on the style of interaction presented and whether they found it enjoyable and/or useful. Eight unpaid British participants volunteered, four female, four male, ages 18 – 72 (mean: 30). Half of the participants were university students and one a retiree. Participants were instructed to "tap the rhythm of the song on the touchscreen, in order to select that piece of music." No limit was made on the length of the queries. The participants were not musicians, otherwise musical background was not controlled. Rhythmic queries were captured using a Nokia N9, running the logging software and variant of the Smith-Waterman algorithm, both developed in C++ using the Qt framework.

Maximum a posteriori (MAP) parameter estimation with a non-informative prior was used in the training, rather than selecting the parameters using the Maximum Likelihood Estimate, retaining some uncertainty so as to mitigate overfitting. The goal was for the target song to always be in the on-screen (top 20) rankings, rather than optimising for the highest rankings.

RESULTS

One measure of interest is how rapidly a user can filter their music collection to recall the desired result on screen (i.e. into the top 20 ranked songs or 6.7% of the collection). Figure 6.9 shows the percentage of queries which resulted in such a ranking, with query performance improving with query duration in seconds. Higher rankings are achieved for all query lengths when using the trained generative model. Queries with lengths of approaching 10 seconds always yielded an on-screen result (in the top 6.7% of the corpus). Up to this point the recognition rate improves with query length, as the additional information is incorporated. It is noteworthy that queries over 10 seconds lead to a rapid fall-off in performance.

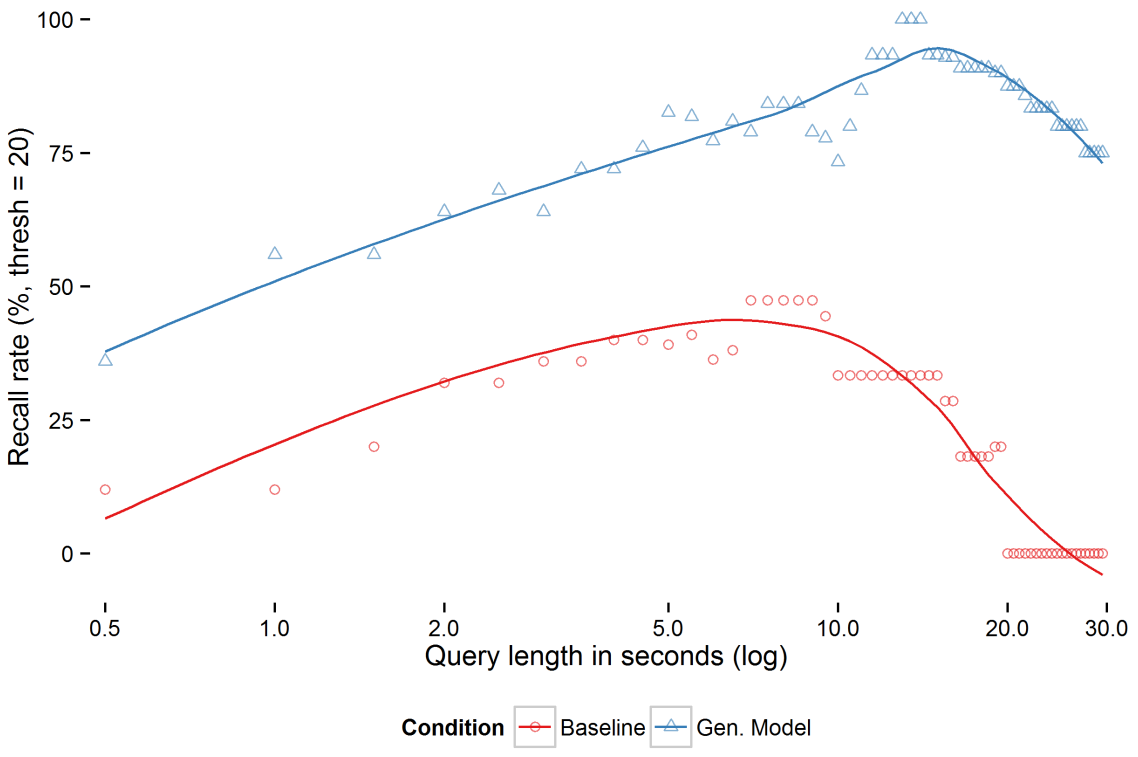

Figure 6.9: Recall rate i.e. percentage of queries yielding a highly ranked result (in the top 20) plotted against query length in seconds. Retrieval performance is shown with the generative model using annotated polyphonic onsets, with monophonic onset detection baseline for context. Performance improves with query duration however falls off rapidly with length.

MEAN RECIPROCAL RANK

Another useful measure is Mean Reciprocal Rank, shown in figure 6.10, reflecting more directly the position of the song within the ranking. The distinction is useful in that, for the purposes of the interaction it is important to get an on-screen result, however a comparison of rank positions gives a more direct indication of the performance of the retrieval.Voorhees (1999) introduced this measure for the evaluation of retrieval systems and discusses its advantages, such as being bounded between 0 and 1. For a set of retrieval results X, with cardinality $|X|$, Mean Reciprocal Rank can be calculated as follows:

$$MRR = \frac{1}{|X|} \sum_{x \in X} \frac{1}{rank(x)}.$$

The much larger gap between the systems in Figure 6.10 than in Figure 6.9 reflects the improvement across the entire distribution of retrieval results. Not only do more retrieval queries return a highly ranked result but the long tail of poor results is also improved.

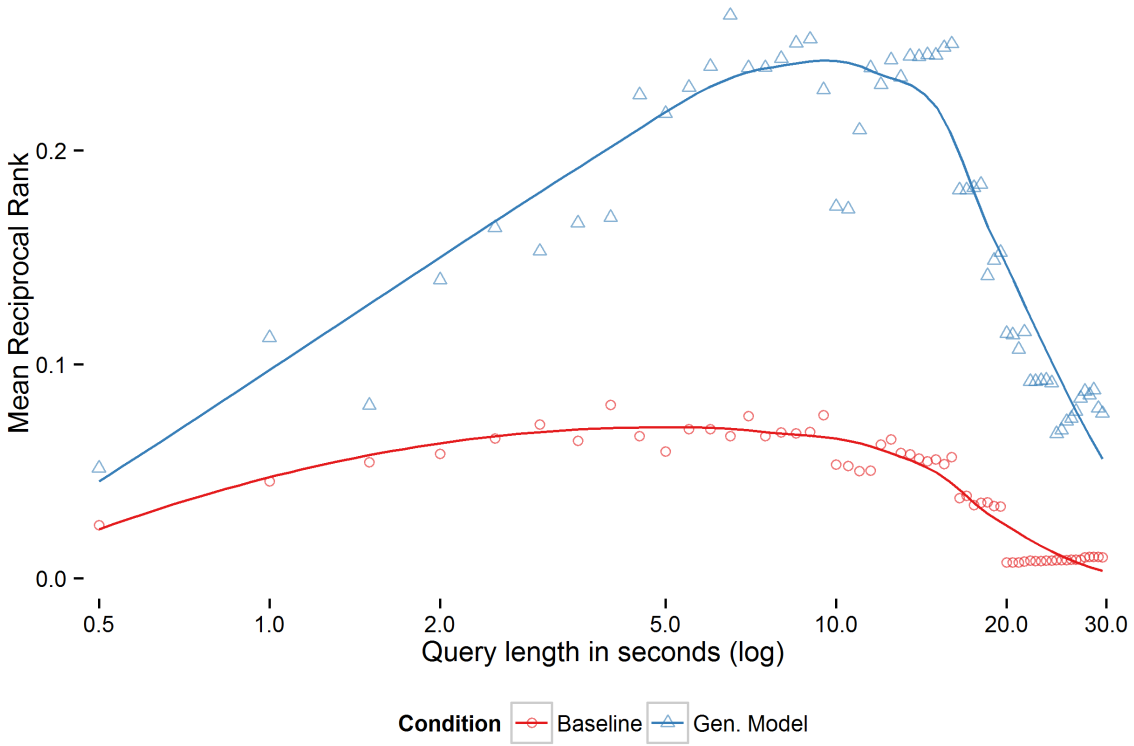

Figure 6.10: Mean Reciprocal Rank (MRR) of rhythmic queries, higher values indicate a better ranking for relevant results. The dramatic gap in performances reflects the improvement across the entire distribution of retrieval ranks, as well as the proportion of highly ranked results.

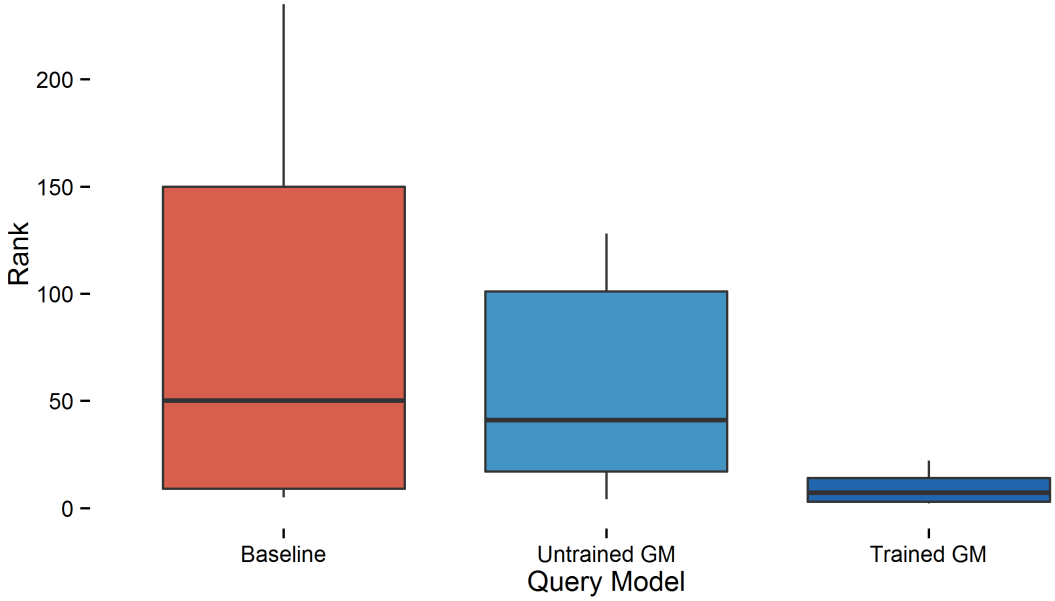

Figure 6.11: Query by Tapping results using the polyphonic generative model, trained and untrained, as well as a baseline using monophonic onset detection

The distributions of retrieval result rankings for each condition are shown in figure 6.11. A paired t-test comparison was performed between the conditions, with Bonferroni corection, showing a statistically significant mean improvement of the trained generative model over the untrained and baseline query likelihood models ($P < .001$). The improvement of the untrained generative model over the baseline was not statistically significant ($P = .279$).

Participants said they enjoyed using *QbT* as an interaction style, often choosing to continue the interaction beyond the requirements of the experiment. The experiment was viewed as a game, with half of the participants requesting further attempts to improve their results. One participant identified in-pocket music selection as an interesting use case, as depicted in figure 6.3. Another expressed concern about the scalability of using rhythmic queries for especially large music libraries. All the participants immediately grasped the concept of *Query by Tapping* and were able to readily produce rhythmic queries.

SINGAPORE FOCUS GROUP

An exploration of the application concept was conducted using a prototype demonstrator with a small focus group of five Singaporean participants – four female, one male. The particpants all used mobile media players and were aged 26 – 59 (mean: 40). They also spoke and listened to music in Chinese and English, giving a view of how the system performs cross-culturally. The demonstrator was presented to them and they were able to interact freely with it, an informal conversation followed to capture their impressions.

Participants were asked to compare the interaction with their usual way of listening to music in a mobile context, to consider the 'in-pocket' use case, whether they would feel comfortable using the interaction style in public and whether the interaction style was suitable for their music. The discussions with participants identified that they all often listened to music using the 'shuffle' feature of their phones or music players, occasionally using the menu interface to select particular music. The use of rhythm to shuffle music was well received as a superior option to random shuffle, with P3 noting that random shuffle can lead to inappropriate music selections at night and that shuffling by tempo could avoid this. Participants were very positive about the in-pocket use case, with P1 saying that it "means I can select music when it's raining" and P2 saying that "it can be hard to take my phone out to change song because I have small pockets." P3 and P4 noted that fear of theft sometimes prevented them from removing their phone to select music. Social acceptability of the interaction is important as the use-case for in-pocket music selection includes being in mobile, public contexts. None of the users had an issue with tapping in public (P5 likened it to drumming on their lap).

Participants also offered up some concerns about the interaction technique. P2 doubted their ability to produce suitable rhythmic queries for all of their music despite achieving good performance with the demo, citing their lack of musical knowledge. They did however appreciate the inclusion of the tempo-only sorting ability to mitigate the need for accurate rhythmic querying. P5 pointed out that the 'intensity' of their taps is an overlooked feature of rhythm and expected it would be interpreted.

6.5 DISCUSSION

The results show that accounting for subjectivity in users' rhythmic queries greatly improves retrieval performance despite increasing the target space. While validity was shown for average music collections, scalability is a concern, with larger music libraries requiring additional sources of evidence for effective retrieval. A further concern is the drop-off in performance for large rhythmic queries, possibly due to the user having a limited section of music in mind when they begin producing a rhythmic query. The results suggest that performance for this and existing techniques may be greater if query length is limited. Further improvements for the generative model should also allow for participants who tap one instrument exclusively.

As expected, participants stated they used shuffle frequently – the ubiquity of shuffle is noted in Quiñones (2007) and attributed to the popularity of the iPod shuffle MP3 player. Similarly, users felt comfortable with the interaction technique – work by Rico and Brewster (2010) has shown that users feel comfortable producing rhythmic and tapping gestures in public. It is of interest that all of the participants felt the interaction suited both their English and Chinese music, existing techniques such as browsing by genre are culturally specific and so this is potentially an advantage of this interaction style.

INCORPORATING SUBJECTIVITY

It could be argued that one would of course expect better results from the use of ground-truth polyphonic music data than from the use of detected onset events. The untrained generative model provides an example of this, showing an apparent though non-significant improvement over the baseline (onset detection). After training the generative model, the improvement in retrieval performance is dramatic. It is not surprising that incorporating knowledge about the user improves the performance of the system. The use of a generative model does not in itself yield much of an improvement however it provides a mechanism by which one can incorporate the prior knowledge about the user. It is only when the issue of subjectivity in how users produce their queries is addressed that significant performance increases are seen. The subjectivity was addressed through the use of a simple model based on initial discussions with users, it is likely that far more powerful models could be constructed.

SCALABILITY

The techniques employed aimed to ensure a top 20 (on-screen) result rather than optimise rank at the risk of failed queries. That users were unsure of their ability to achieve this level of performance indicates that further work could be done to improve users' confidence during the interaction, for example with real-time feedback. It is to be expected that as the size of the music collection is increased, retrieval performance using rhythmic queries will fall. The results show this style of interaction to be valid for an average music collection of 300 songs. Performance is far greater as collection size is reduced, with queries yielding first ranked results 65% of the time when the collection is halved to 150 songs. The Bayesian approach allows for this issue of scalability to be addressed through the introduction of additional sources of evidence. For example a different interaction style could use sung queries – providing pitch, rhythm and tatum as evidence. Such an interaction would also benefit from the user query modelling approach introduced here. Other evidence sources could include the dynamics of the tap events, for example a 'strumming' action could denote guitar events.

Rather than implement a simple retrieval of a musical work by tapping its rhythm, the user's entire collection of music is re-ordered according to the rhythm and tempo evidence. This means that for this collection or larger ones, the drop-off in retrieval performance is more acceptable as the user is still able to assert control over their music collection, ordering it according to the evidence they provide. The queried music could even just be a subset of the music – landmark tracks, which the user selects to indicate the type of music they want. At the very least, the user can shuffle their music by tempo and rhythmic similarity. The intended song need only be ranked in the top 20 results displayed on-screen, with the user then able to select the song easily.

LIMITATIONS

While the benefits of the techniques presented have been demonstrated, much work remains in studying rhythmic querying as an interaction technique. The evaluations here aimed to validate the use of the generative model. A wider study could provide further insight into rhythmic querying behaviour – for example, exploring the suggested additional features to incorporate tap intensity, single-instrument annotation and song-specific tapping strategies. A key limitation of this work is that the training and validation queries are acquired in the same session - a longitudinal study may show that users' rhythmic querying behaviour changes over time. Another potential limitation is that, prior to the evaluations, participants were briefed as to the nature of the experiment. It is possible that participants then selected songs to query which they felt confident of being able to tap to – introducing a potential source of bias. During the experiment, participants generally focused on simply finding a familiar song to query and enjoyed the 'challenge' of the task, so it is unlikely that this 'tap-ability' bias had much effect. Overall, it has been shown that *QbT* can be greatly improved with the use of a trained generative model and is an interaction technique worthy of much further exploration.

QUERY LENGTH

A surprising result is the sharp drop-off in retrieval performance with query length. One might expect that as more evidence is introduced, retrieval performance would increase, indeed such a relationship is seen in queries up to $10s$ in length. It is unlikely that users recall the entirety of a song in advance of producing a rhythmic query – instead they would select a memorable or distinctive passage of the music. The drop-off in performance could reflect that users have continued beyond the salient part of the music they had in mind. This issue could be addressed by limiting the length of rhythmic queries or by providing real-time visual or haptic feedback to the user so that they entrain with the song as it begins playing. This result has wider implications for rhythmic interaction (for example in the use of rhythmic 'hot-keys' in Ghomi et al. (2012)) in that it indicates an upper length for rhythmic patterns. More generally, this result could suggest that any music content based querying technique such as humming or singing may also suffer from falling performance for queries over ten seconds. It is worth noting that while mean rank result improves with the use of the generative model and with training, the most significant change is in the 'long tail' of poor results. This work not only improves mean query performance but also makes for a more consistent user experience, cutting off the long tail of poor results caused by subjectivity in query production.

FURTHER CHALLENGES

Having identified the benefits of improving consensus with the user with a trained model, there is an opportunity for further study in using feedback as part of a consensus-building interaction with the user. The rhythmic matching algorithm could be improved further through the use of music theory, for example where a whole note is replaced by four quarter notes, the penalty should be very little. Such techniques have been applied in similar efforts in matching pitch in melodies, leading to the Mongeau and Sankoff (1990) algorithm and eventually to services such as SoundHound. Such work, combined with improvements in onset detection, could allow robust commercial applications of rhythmic music retrieval. Given that users can recall a great number of musical works and the results shown here, future research could build upon this work to use musical rhythm for interaction tasks other than just the retrieval of music, e.g. tapping a rhythm on a phone in-pocket to dial a corresponding contact when using a bluetooth headset.

This work has implications for future research in interaction with music, demonstrating a need for considering user variance. In the following chapter, generative models of user queries are again used for inferring user intent, coupled with prior knowledge of their interest in a music space.

THESIS DEVELOPMENT

The layout of this thesis, from defining and measuring engagement through to designing for and adapting to it, is structured to provide a coherent research narrative. It is not entirely reflective of the order in which the research was conducted. The investigations in this chapter were performed early on in the work of this thesis, providing a platform for the development of many of the ideas presented. In particular, the use of the mobile, casual *QbT* interaction to retrieve specific songs is incongruous with the concept of a casual listener having less control as discussed in section 3.6. The limited attention and engagement available for mobile interaction helped to motivate the consideration of music engagement. This chapter also explored the use of queries based on tempo, as a more casual retrieval technique. The use of a single interface that spans multiple styles of retrieval engagement is further explored in chapter 7.

6.6 CONCLUSIONS

Previous efforts at *QbT* were improved upon by using trained user query models to sample from polyphonic music data. Using a generative model of subjective queries has taken rhythmic music retrieval from a concept with potential to a usable interaction style. These user models allowed for a dramatic improvement in retrieval performance, with the intended song always appearing in the top ranked results. This work highlighted the issues caused by subjective music queries and developed a personalisable music retrieval system. Users enjoyed using the system in trials, often asking to continue use beyond the experimental requirements in order to attempt to improve their ranking. Users were able to generate rhythmic queries from their subjective interpretation and memory of music rather than using a memorised rhythmic pattern. Removing the need for memorisation in this way has applications beyond music retrieval, for example the work on rhythmic hot-keys could benefit from the presented approach.

The interaction technique presented in this work enables users to enjoy a casual style of music retrieval – empowering them to interact with their music in new contexts such as in-pocket music selection. By shuffling the music playlist by the rhythmic query, users can provide uncertain queries about a type of song or query for a particular song without having to recall its title etc. The techniques developed here achieve the goal of casual in-pocket music control with users able to see the benefit of the interaction style and identify use cases relevant to them. The personalised rhythmic filtering presented offers a music retrieval style designed to support both low *interaction engagement* and low *retrieval control* re-ordering by tempo, or for users to engage more and query for a specific song's rhythm.

7. ADAPTING MUSIC RETRIEVAL TO USER ENGAGEMENT

ADAPTING a music retrieval interaction to the user's desired level of effort and control would enable a single interface to span the range of user engagement. The hypothesis explored in this chapter is that a recommender system could relieve users of the burden of control according to the level of cognitive effort they are willing to invest. Users would be able to seamlessly transition from a casual style of interaction akin to using a radio to more controlling styles such as specifying a particular sub-area of interest in a music space, or even selecting individual songs. An example of such adaptation can be seen in Pandora's recommender system,[1] though this interaction is again at a single level of engagement (a simple like or dislike rating of tracks). A demonstrator system is developed, with a mood-based music retrieval that allows users to select a level of control over a recommender system. An agent-based evaluation characterises this interface, and how it can span levels of user engagement, with additional comments elicited from participants in a design session.

[1] `http://www.pandora.com` (11/08/14)

The tablet-based demonstrator system shows how users can seamlessly take control over recommendation, from hearing any popular music, through to more specific sub-regions of music, all the way to the user explicitly selecting individual items. Users are able to explicitly denote their engagement with a pressure sensor mounted on the tablet's bezel – employing a metaphor of physically engaging. A large music collection is arranged into a coherent one-dimensional space (ordered by mood), such that broad selections of a mood of music can be made. The system employs a generative model to predict the user's input for each available song. When no control is exerted the expected input is broad and uncertain, with the system using prior evidence (such as song popularity) to make an intelligent inference of which music to recommend. As the user exerts control, the expected inputs become more specific, allowing the system to infer the user's intended song.

There are many sources of evidence of users' engagement context that could be considered for adapting music retrieval. Music engagement is explored in chapter 3, with measures of user engagement and listening behaviour developed in chapter 4. While these give some evidence for adapting to the user, there has also been work on measuring engagement in the current listening context, for example by detecting rhythmic movements (Kuhn et al., 2011). As a proxy for such methods, in order to explore this design space, the system developed in this chapter allows users to denote their engagement explicitly by pressing a pressure sensor – employing a design metaphor of physical exertion as engagement. The development and evaluation of this sensor is detailed in Appendix A.

Engagement & Control

Chapter 3 has explored user engagement and interactions that bridge from casual to engaged styles of use. If a music retrieval system enabled users to set their level of engagement, it could cover the maximising-satisficing scale – catering to individual users and their music-listening contexts. This dynamic has important implications for the user's sense of ownership of the music interaction. It is likely that users would no longer feel the sense of pride or embarrassment related to a track selection (as described in Cunningham et al. (2004)) if control was shared with an intelligent music system.

In casual interactions, where the user exerts less control, a system can act more autonomously – making inferences from prior evidence about what the user intended. This way of handing over control was termed the 'H-metaphor' by Flemisch et al. (2003), where it was likened to riding a horse – as the rider loosens the reins, and exerts less control, the horse behaves more autonomously. Music retrieval would benefit from spanning the range of *retrieval control* in this fashion, allowing users to set a level of engagement appropriate to their context. Users would have a corresponding degree of control over the recommendations, ranging from biasing them to a style of music, through to the explicit selection of a particular track.

INFERRING LISTENING INTENT

A music retrieval system could adapt to the user's engagement by acting more autonomously when the user wishes to have less *retrieval control*, i.e. when they do not have a specific goal in mind. This can be described in terms of the information-theoretic viewpoint discussed in chapter 3 – as the user provides fewer bits of information to reduce the uncertainty about which song to play, the system must increasingly provide the necessary information. The input from the user can thus be considered coarser as they become less engaged, having less information content and thus less weight on the behaviour of the system, instead indicating the 'broad strokes' of the user's intent. Their intended selection can then be inferred using prior evidence of music-listening intent, for example the user's previous listening history, social recommendations or the overall popularity levels of each song.

While the input from the user may be coarser and less precise when the user is less engaged,[2] in other cases similar input must be evaluated in the context of different engagement levels. For example, a touchscreen event when a user is casually browsing through an overview of a music library should be interpreted more coarsely than an explicit selection of a piece of music. In order to correctly interpret this input, the system must be aware of the engagement context – either by sensing it (e.g. with accelerometers (Kuhn et al., 2011)) or by it being made explicit by the user. The handover of control may be continuous or occur on a number of discrete levels, for example moving from proxemic to touch interaction. This discretisation allows distinct operating modes which are clearly delineated however burdens users with managing these multiple modes. As an alternative, the exploration of engagement in this chapter treats it as a continuous variable, acting as a parameter in the inferential system. This continuity allows for a seamless, consistent and natural interaction.

[2] e.g. Pohl and Murray-Smith's 2013 work on casual interaction

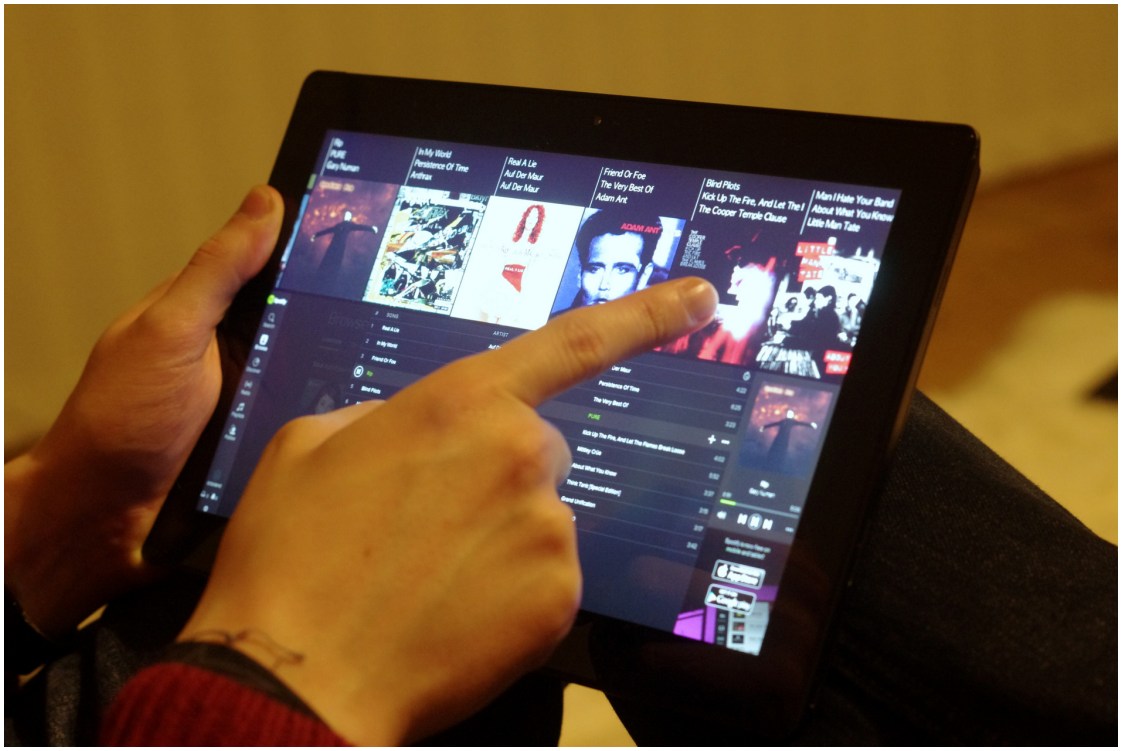

Figure 7.1: The exemplar system (above) adapts to the user's engagement (their desired level of *retrieval control*), allowing users to explore music at a high-level (e.g. selecting a broad mood such as angry music) or seamlessly engage, zooming in to make specific selections (e.g. albums in a more specific area). Users retain the ability to fully engage and control the system into an exact album selection.

7.1 EXEMPLAR SYSTEM

This chapter sees the augmentation of a popular music retrieval interface (Spotify[3]) to explore and demonstrate how music retrieval can be adapted to user engagement. This was developed as a tablet-based prototype, with a pressure sensor attached to the bezel, as in Figure 7.1. Also, a semantic zooming view of a simple linear music space was added, enabling both casual and engaged forms of interaction – giving users varying degrees of control over the selection of music. As users engage, applying pressure, their input increasingly influences the music recommendation, up to taking complete control. The music was arranged on one axis as this is the simplest – offering a mode with the least possible engagement (Figure 7.2). By allowing users to make selections from the *general* to the *specific*, the new interface supports both maximising and satisficing and spans levels of engagement, or more specifically, the possible levels of *retrieval control*. Users can make broad and uncertain *general* selections to casually describe what they want to listen to. However, they can also exert more control over the system and force it to play a *specific* album (Figure 7.3). An online overview video of the system and the interaction technique is available.[4]

[3] `http://www.spotify.com`
[4] `http://www.dannyboland.com/mobileHCI15/` (20/02/15)

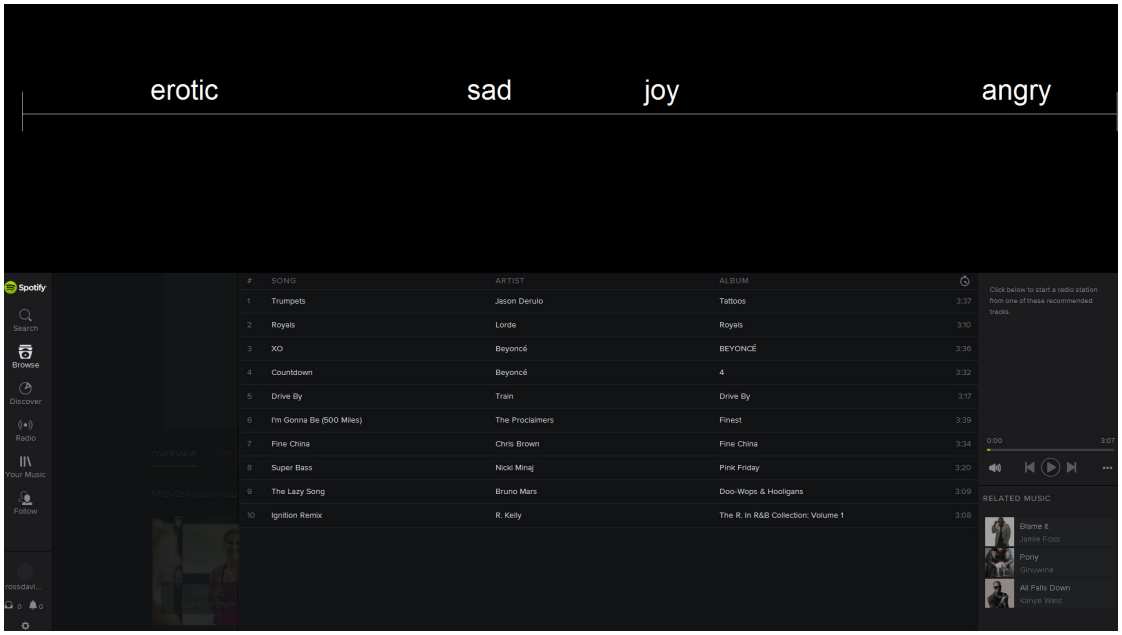

Figure 7.2: The zoomed out (low engagement) state allows users to casually make a selection at any point in the mood space. Such a selection has a high uncertainty with regard to individual albums and so the recommender agent uses popularity as prior belief of listening intent.

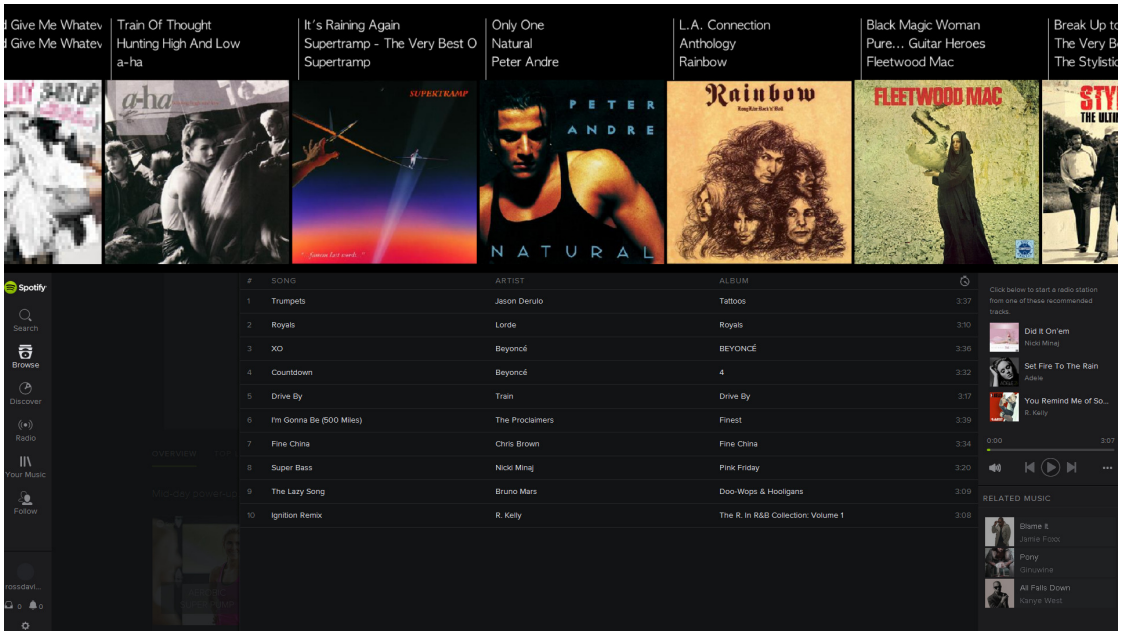

Figure 7.3: The zoomed in (high engagement) state allows users to make a specific selection of an album, having zoomed in to a specific area of the mood space. Such a selection has a low uncertainty with regard to individual albums and so the recommender agent has negligible influence.

INFERENCE

The handover from casual to engaged interaction relies upon an inferential model of user listening intent. It is assumed that when making casual selections, users will navigate to a broad area of musical interest, e.g. happy music. It is also assumed that as users engage and wish to make more specific selections, they will use the richer album art and metadata feedback to navigate to a few (or one) albums of interest. These assumptions about the user's behaviour are encoded as a generative model, i.e. one which predicts the user's inputs for a given level of engagement, as in Figure 7.4. Put simply, it predicts casual users will point roughly near the type of music they want and engaged users will navigate to exactly what they wish to hear. This generative model predicts user input i_x along the x axis conditioned upon a target song and level of engagement. The input is modelled using a Gaussian Mixture Model, with a component corresponding to each song s_i in the music collection (omitting distant songs for scalability). Each component is a Gaussian distribution, with the centroid positioned according to the song's position x_s in the one-dimensional music projection, with distribution width set by the precision τ parameter:

$$p\left(i_x \,|\, s_i\right) = \sqrt{\frac{\tau}{2\pi}}\, e^{\frac{-\tau(i_x - x_s)^2}{2}}.$$

The precision τ parameter is inversely proportional to the unit variance σ^2 of the distribution. As users apply pressure, the precision is scaled by the denoted engagement $E \in (0, 1]$:

$$\tau = \frac{E}{\sigma^2}.$$

This generative model can be used to infer a belief that a given song s_i is of interest to the user, conditioned upon an input position i_x and the level of user engagement E:

$$p\left(s_i \,|\, i_x, E\right) = \frac{p\left(i_x \,|\, s_i, E\right) p\left(s_i \,|\, E\right)}{p\left(i_x \,|\, E\right)}.$$

The prior belief $p\left(s_i \,|\, E\right)$ over the music space allows existing evidence of listening intent to be incorporated, conditioned on the current level of engagement. This work uses the music's current popularity on Spotify so that the system can recommend popular music to the user. A simplifying assumption is made that only expected input varies with engagement and not music preference, though a more informative prior may be possible. The system's behaviour is characterised using an agent-based approach, enumerating the possible inputs. As shown in Figure 7.6, the less engaged the user is, the more the system relies upon its prior evidence – selecting popular music in the broad area navigated to. At higher levels of engagement, the system defers to the user's input, selecting nearby songs except those it considers unlikely.

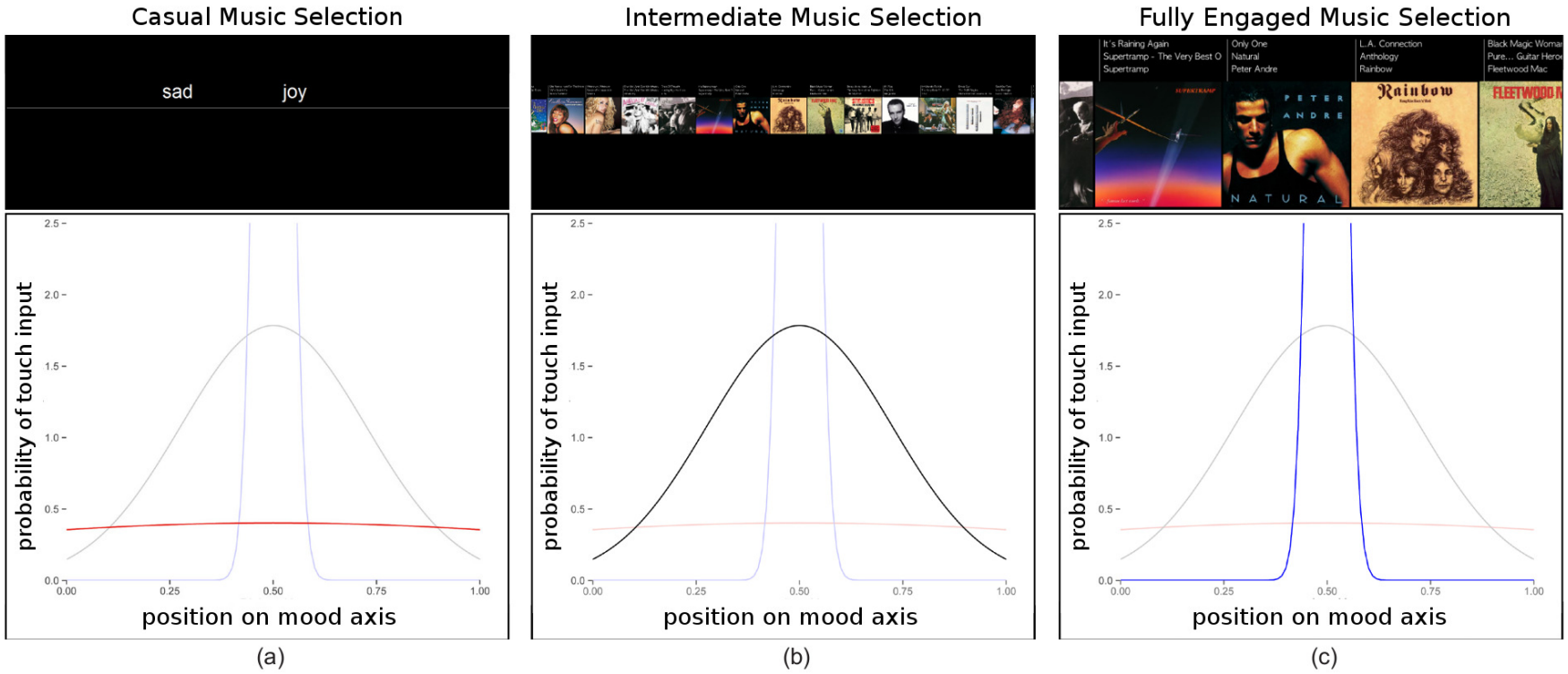

Figure 7.4: As the user exerts control, the distribution of predicted input for a given song becomes narrower. This means the system can give more weight to the user's input and infer a belief about their listening intent over fewer songs. (a) shows a casual music selection, with no pressure applied, the user sees the entire axis with mood labels to give a sense of the layout. (b) shows a more engaged music selection, with the user now applying pressure, the view zooms semantically to show album artwork for a range of music of a particular mood, allowing users to get more specific recommendations of popular music. (c) shows a fully engaged music selection, when maximum pressure is applied, users are able to zoom in to view a specific album artwork and make an exact selection, with the system deferring entirely to their input.

EMERGENT BEHAVIOUR

When the user has not applied pressure, they have a low level of engagement in the retrieval. The view of the music space is thus zoomed out, with mood labels describing its layout (Figure 7.4a). The inferred music selections are broad, covering an entire region of their collection and are biased by prior evidence (popular tracks). At low levels of engagement it is likely that most tracks played would be highly popular tracks. This behaviour is a design assumption. Users may want the system to use other prior evidence, such as recency of album release, as popularity may limit access to the long tail of music (see section 2.3).

When the user applies pressure, the system interprets this as the user taking control to make a more exact selection. The inferred selection is more specific, and the view zooms semantically to show the album art of the selection (Figure 7.4b). This selection is a combination of evidence from the user's navigated position with prior evidence, i.e. song popularity. Users retain the ability to make explicit selections by fully applying pressure (Figure 7.4c). By varying the pressure, users seamlessly set their level of engagement and control over music recommendation and selection.

CONDITIONAL DYNAMICS & SEMANTIC ZOOMING

The level of detail presented to the user should be appropriate for their level of engagement. This is a natural application for semantic zooming – users making broad selections would not benefit from seeing the artwork of each album and so the view zooms out, showing only general labels such as a mood. This approach can be generalised, to condition other aspects of the interaction upon the engagement and inference of user listening intent. Not only can songs be played that are considered of interest to the user but the mappings of the input and output modalities could vary according to the engagement and inferred selection. An example would be adjusting the semantic zooming to fit the posterior (inferred) distribution, ensuring all songs of interest are displayed. An appealing benefit of such conditional mapping functions is that the nature of the inference is exposed to the user. When swiping through the music space, users could find the navigation attracted to or repelled by songs according to popularity. This stickiness could also occur in zooming, with extra pressure required to zoom into unpopular tracks. Such sticky interface elements echo past work in HCI such as Cockburn and Firth (2004), where they were shown to be beneficial to target acquisition and popular with users. The greatest disadvantage to generative modelling is that where the model does not conform well to the user, the user will be unable to predict the behaviour of the system. As the interaction is increasingly conditioned upon the inference, the exposure to this problem is greater.

Shneiderman and Plaisant (2004) recommend adhering to the principle of 'direct manipulation' such that the mapping between input and output is clear to the user. In this work, a conservative approach is taken, engagement is mapped linearly with applied pressure and the semantic zooming is done linearly with engagement. Only the graphical feedback is manipulated to display the posterior distribution. Tracks are sampled from the posterior in real-time and their album artwork is highlighted with a faint glow. Tracks are also shown in the playlist frame of the Spotify UI in order of inferred listening intent.

A SCALABLE MUSIC SPACE

The simplest possible representation of a music space is a one dimensional projection. An assumption of the system is that similar songs are placed near to each other, such that when users are disengaged, they can select a whole coherent region of the music space. Information about the mood of the music is used to arrange the collection however genre or other features and metadata could be used. A high dimensional music feature space was produced using MoodAgent,[5] a commercial music signal processing system focused on mood-related music features. While a single one of these features (e.g. tempo) could be used to arrange the music collection along one dimension, this work attempts to maintain more of the information by using non-linear dimensionality reduction.

Venna et al. (2010) introduce an information retrieval perspective on such projection techniques, presenting their state-of-the-art NeRV algorithm for balancing recall and precision in the projection. This trade-off between keeping similar items together and dissimilar items apart can be modified using a λ parameter. NeRV is used here to project the high dimensional music feature space into a one dimensional space – Figure 7.5 shows how the original features influence this projection. In this example, anger is inversely correlated with tenderness, with joyful music peaking in the centre of the projection.

The resulting one dimensional projection broadly reflects the culturally universal arousal aspect of mood (Egermann et al., 2015), spanning from quiet, calm music to faster, aggressive music. The use of a single dimension is justified by the observation by Egermann et al. that the valence dimension of mood is subjective. The relatively denotative arousal aspect of mood is more readily captured by machine listening methods. The use of mood itself for music retrieval is supported by evidence that mood features describe music as well as genre (Boland and Murray-Smith, 2014) and that users frequently listen to music according to mood (Lonsdale and North, 2011).

[5] `http://www.moodagent.com/` (11/08/14)

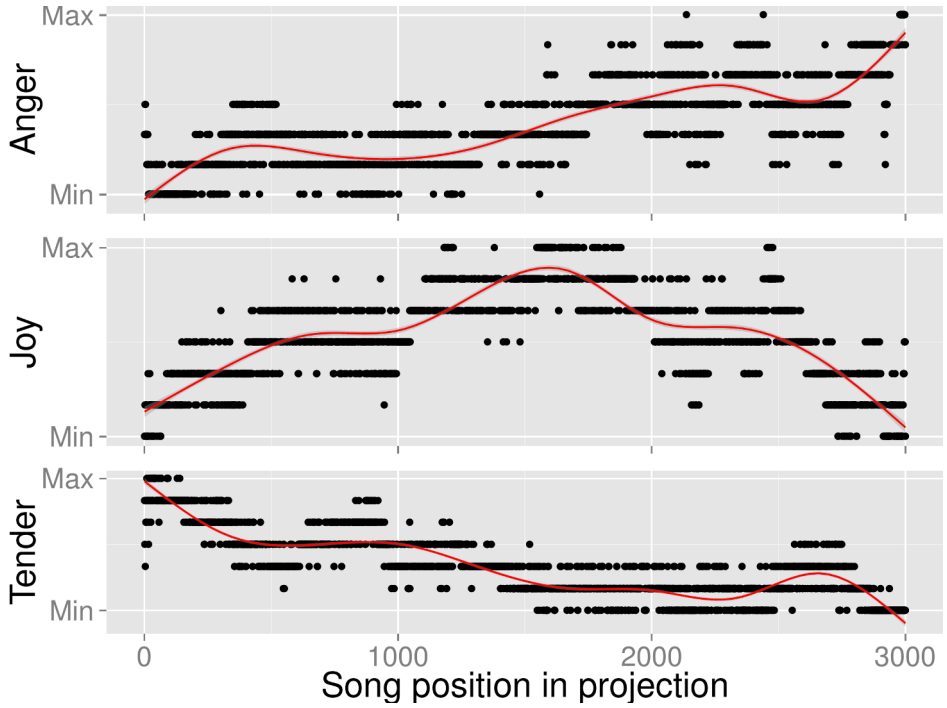

Figure 7.5: The music is arranged by taking the mood features such as Anger, Joy and Tender and creating a one-dimensional projection. Users are then able to make broad selections from a mood region, as well as pick individual albums.

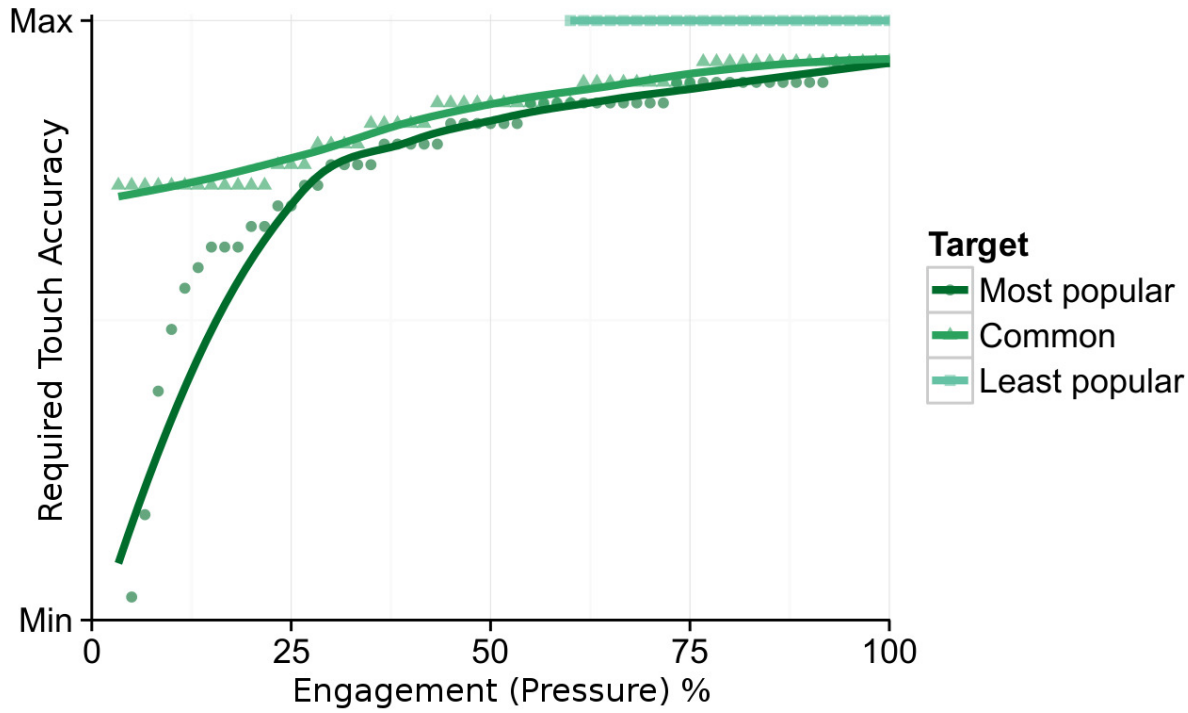

Figure 7.6: Agent-based characterisation of the required accuracy for song selection against user engagement. Popular songs can be selected by casually navigating near them. Unpopular songs are only selected when the user is highly engaged and navigates exactly to them. As the user applies pressure, denoting increased engagement and their wanting greater *retrieval control*, they take control from the recommender system. As the user takes control, the songs become equally targetable, with the popularity-based prior having diminished influence.

7.2 EVALUATION & DISCUSSION

It is inherently difficult to find an objective means of evaluating such a personalised music system. Schedl and Flexer (2012) identify this challenge as a key reason that little research work has been done in personalised music systems. They note the need to consider the user from an early point in the development process. The use of generative modelling is an example of such a user-centred development approach. It allows the simulation of user input to characterise the behaviour of the system across the space of possible input, as in Figure 7.6. A series of exploratory design sessions were also conducted, to obtain qualitative feedback about the system. The music space used was taken from the UK Top 40 singles charts over the past 50 years. Audio was streamed from Spotify, with the engagement-dependent interface integrated within the Spotify web player UI.

ENGAGEMENT-DEPENDENT RETRIEVAL

As the system is based upon a generative model of user input, it can be used to characterise the input required from users in various scenarios. Using the design assumption of users casually choosing popular music, it can be shown how input required from users varies with the popularity of the retrieved track. Figure 7.6 shows how accurately the user must navigate to a song for it to be played, at different levels of engagement and for songs of different popularity. The behaviour of the system can thus be linked to the music listening profiles identified as targets in section 4.4.

Casual This profile called for allowing users to satisfice and make corrective actions with little effort. The example system supports this with recommendations of popular music from the broad area navigated to, requiring only swipe gestures to select or change music. The recommender system largely controls the music selection, with coarse high-level influence from user input. Notably, it is impossible to select the least popular songs at low engagement.

Engaged This profile called for allowing users to invest effort in making an exact selection of an album. The system supports this by letting users apply pressure to zoom into a specific mood region, or even an album of interest. At high levels of engagement, the user input dwarfs the recommender's influence, with popularity having little effect.

Mixed The intermediate levels of engagement are supported by letting users quickly set their engagement, with the continuous handover of control using the pressure sensor. As users engage, their input has increasing influence. Semantic zooming gives users feedback about how specific a query at a given level of engagement would be.

DESIGN SESSIONS

Six participants, in groups of 2, took part in informal design sessions which focused on discussing the music interaction described in this work, and in particular the benefits of being able to engage with a mood-based, semantically zooming view. Each was given time to freely explore the collection and to experiment with making a variety of music selections at different levels of engagement. This experience was used to generate discussion with each of them about how they currently find and listen to music and how the system might support or enhance those habits. As mentioned in the Collaborations statement, the experimental system was given to Ross McLachlan who then conducted and annotated the discussions with users, which are explored here.

SHUFFLE AND CASUAL MUSIC SELECTION

Five of the participants reported using both shuffle and explicit selection to play music in their regular listening, while only one exclusively used explicit selection. Participants stated that they typically use shuffle for background music or alleviating the social pressures of selecting music, as discussed in section 3.1. This supports the prediction that the ability to vary the degree of engagement is desirable, with four participants often creating playlists of songs specifically to shuffle over. These playlists were based on personal heuristics or 'idiosyncratic genres' as described by Cunningham et al. (2004). They were often kept to a small number of songs to guarantee similarity between items. It is worth highlighting that users are seeking to engage with their music system at a level between shuffle and explicit selection, which is largely unsupported in existing retrieval interfaces.

CURATION AND TAKING CONTROL

Participants all expressed a desire to curate the collection. One suggested that there could be more room for subtle differences between tracks in a personal collection, which would allow finer grained control over even uncertain selections. One pair perceived the underlying structure of the music space to be close to random and felt that having more control over the music in the collection would lead to a greater understanding of the organisation, commenting that "I'd like to be able to apply more of my own decision making." This sentiment was echoed in comments about the labels. One participant commented that self-defined labels would be better, even if not fully representative of the underlying mood – "people might be annoyed if all their favourite albums are classed as 'sad'... I would prefer 'melancholic', it has nicer connotations." Clearly, a desire for customisation is present that could be accommodated for by allowing users to define the underlying collections, as opposed to generating it from the Top 40 charts.

SELECTION FEEDBACK

The visual feedback used for low engagement (mood labels on the axis) misled users into assuming that the music was classified into distinct mood categories. Users did not perceive the probabilistic nature of the mood space. For example, the point where 'Joy' ended and 'Anger' began was a constant topic of discussion. There is an inherent uncertainty to the music layout and selection that was not conveyed. More abstract feedback would have better communicated the nature of the selection, and that each song is a mixture of moods. For example colouring the music space according to mood, with a blending of colours as mood overlaps, however any extra complexity would have required more user training and reduced the system's ability to support the most casual retrieval scenarios.

Participants found the semantic zooming to be useful, but felt that the granularity should increase more evenly. Making very broad or very precise selections was easier than making intermediate selections. The visual feedback presented in intermediate positions, the line of very small album artworks, "invites you to zoom in" rather than providing feedback about what selection will happen at that zoom level. Participants wanted even finer control over their level of engagement in their music retrieval, which lends support to the argument for engagement-dependent music retrieval.

SOURCES OF EVIDENCE

The present implementation of the system uses music popularity as a source of evidence for weighting songs as being more likely to be played. This comes from a design assumption that satisficing is supported by recommendations based on a prior such as popularity. The design session revealed that users felt that in doing this, recent additions to their music collection would be underplayed. By including when the album was added to the collection as prior evidence, the selection can be biased towards recent additions as well and this use case can be supported. A wide range of other sources of evidence could be used in place of popularity, such as the user's music listening history or, from the commercial perspective, the artists currently being promoted by music labels.

MUSIC SPACES

The behaviour of this system is closely linked to the music space used. In this work the music space used is likely to be familiar to all participants, however this does introduce some issues. In early testing the music space comprised the most popular 5000 tracks from the UK charts over the last five decades, it was found that the resulting music space was too homogeneous to navigate meaningfully. In an effort to increase the diversity of the music space, tracks were randomly sampled from the music collection instead, yielding the distinct mood regions seen.

Users disliked when dissimilar songs were together. A benefit of using the NeRV algorithm to project the music arrangement is being able to adjust the λ parameter to favour precision (keeping dissimilar items apart) over recall (keeping similar items together). Using feedback from users while iterating the design of the system, λ was biased strongly toward precision to keep dissimilar items apart and improve perceived system quality. The need to avoid stark outliers has been noted within MIR; Paul Lamere of Echonest introduced the 'WTF test' for automatic playlists, with systems scored negatively for each outlier.[6]

7.3 CONCLUSIONS

Having motivated the need for music retrieval that adapts to user engagement throughout this thesis, this chapter shows one way in which this may be achieved. The exemplar system serves to illustrate not only the benefits of this adaptation, but also some of the inherent limitations. Allowing users to seamlessly change their engagement, from scanning broad labels to specific artwork, supports distinct use cases. It is less clear, however, how to support intermediate states of engagement. Additional levels of feedback can be added, such as labelling landmark tracks or artists, or displaying other metadata. There is an opportunity for further work on providing overviews of music collections, in particular with semantic zooming. The use of audio feedback, for example by mixing samples from the inferred track selections, is one such approach.

MOOD

Information about the mood of the music is used to arrange the collection, however, genre or other features and metadata could be used instead. The choice of mood for use in this work is supported by studies of why users listen to music, such as Lonsdale and North (2011). The music features adopted are widely used, having been acquired from a popular commercial service. They provide a coherent music space however this is only to serve as an illustration. Reliably detecting mood and genre from audio samples is an ongoing area of research, with a number of challenges remaining (see chapter 2 and Sturm (2014a)).

Reducing the music features to one dimension provided the simplest music space for users to navigate, at the lowest levels of engagement. The more complex the music space presented to the user, the more time would be required for familiarisation with the space. Future work would benefit, however, from considering a music space with higher dimensionality, as this would support an investigation into more highly engaged exploration.

[6] `http://musicmachinery.com/2011/05/14/how-good-is-googles-instant-mix/` (11/04/14)

MARKET-BASED EVALUATION

While the behaviour of the system has been demonstrated using technical measures and qualitative feedback, it would be desirable to study the longitudinal impact that using such a system would have on listening behaviour. At this point, rather than conducting further small-scale studies, it is preferable to widely deploy a full music system, which users can adopt for their day to day listening. The concepts presented in this work have been applied in a commercial, tablet format, engagement-dependent music retrieval system – the BeoSound Moment, announced at CES 2015. This product is part of an ongoing commercial evaluation, with some initial evaluation efforts described in chapter 8.

Part IV

Outlook & Beyond Music

8. Industrial Application

Industry adoption of concepts and methodologies developed in academia can provide a platform for evaluating work 'in the wild', implemented to a standard comparable to users' existing products. The development of Bang & Olufsen's BeoSound Moment was informed by the engagement-dependent, mood-based music retrieval interactions discussed in this thesis. It incorporates an interface that allows users to determine their engagement with a music retrieval interaction, and to choose how much control to have over their music recommendation. It also presents the music retrieval using a low-dimensional projection of a music space, based on the mood features. As part of the product development process, the metrics developed in chapters 4 and 5 were used to capture changes in users' music listening behaviour once they adopted the BeoSound Moment as their retrieval device. This chapter documents the exploratory evaluation of this product, and the impact of the ideas discussed in this thesis.

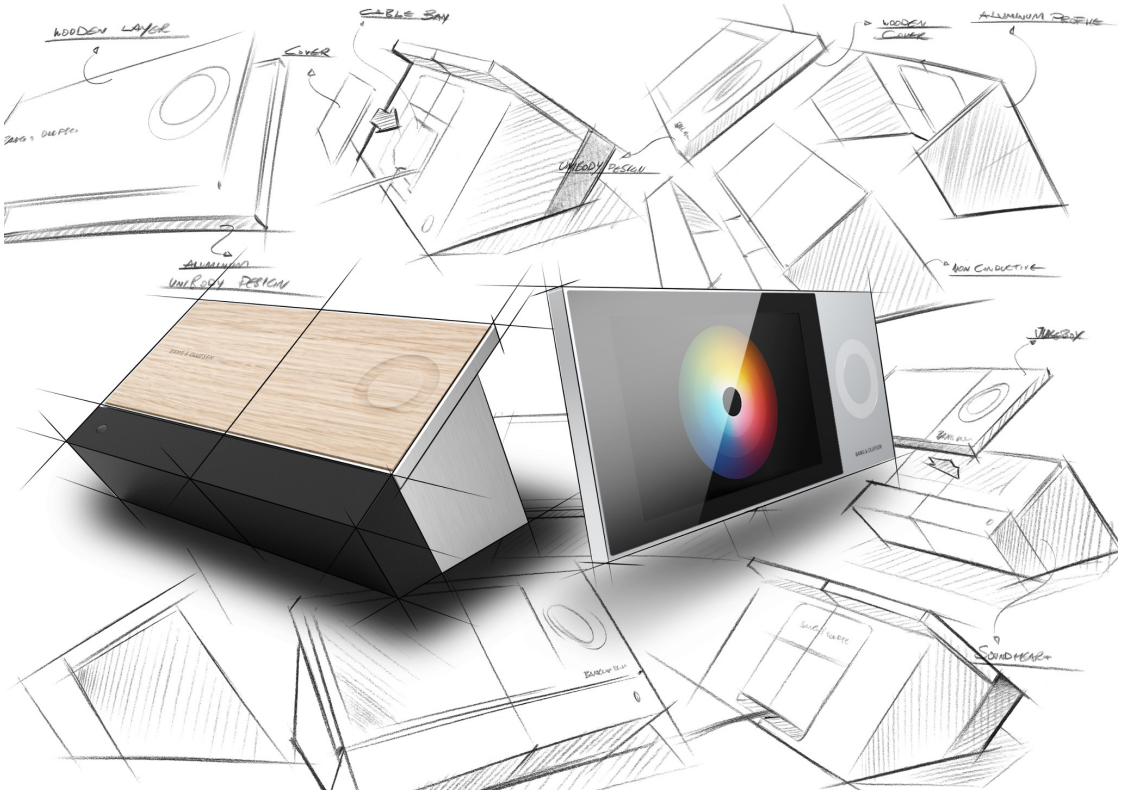

Figure 8.1: The development of the BeoSound Moment was informed by the engagement-dependent, mood-based music retrieval interactions discussed in chapter 7. Press materials, ©Bang & Olufsen.

8.1 BEOSOUND MOMENT

The BeoSound Moment is a commercial implementation of an engagement-dependent music retrieval system, depicted in figure 8.1. A mood-based presentation of the available music space is used, similar to that in chapter 7. As well as having alternative music selection options (genre, radio, album, etc.), the Moment differs from the work in this thesis in that engagement is separated into three distinct modes: *core*, *familiar* and *discovery*. These modes control the balance between the music selection being based on the user's music library and recommendations from a cloud music service. Users are able to listen to their own music by making a mood selection at the *core* level of engagement. Strongly seeded recommendations are made for mood selections at the *familiar* level. Users can hand over control, receiving broad recommendations by making mood selections at the *discovery* level.

While some of the results of testing during development are discussed here, the product has been publicly announced at the Consumer Electronics Show 2015[1] and will soon be evaluated 'in the wild' by customers.

[1] `http://bogone.blob.core.windows.net/static/files/press/BeoSound_Moment_Press_release.pdf` (06/01/15)

Figure 8.2: Music is arranged around a 'MoodWheel' that is coloured according to the mood of the represented music. Users can select moods from the wheel, with the radius of their selection indicating their chosen level of engagement. Press materials, ©Bang & Olufsen.

MOODWHEEL

The engagement-dependent selection mechanic is presented to the user as a 'MoodWheel' (figure 8.2) - the available music is arranged around a wheel, which is coloured according to the mood of the represented music. The wheel comprises three distinct layers that correspond to the *core*, *familiar* and *discovery* styles of selection or recommendation. Users can touch a position on the ring to hear music of the selected mood from their own collection, or recommendation, according to the selected engagement level.

In contrast to the similar interface in chapter 7, the range of engagement levels is discrete rather than continuous. While this allows users to more easily select and return to stable states in their interaction, it does require users to maintain and switch between parallel mental models of retrieval behaviour. By limiting the number of engagement levels to three, the risk of users being confused is mitigated, though at the cost of users being able to finely control the handover between their music and recommendation.

8.2 EVALUATING USERS' CHOICE OF ENGAGEMENT

The engagement-dependent approach developed in this thesis is based upon the motivation that users would like to perform their music retrieval at a range of engagement levels. Being designed to afford such an interaction, the BeoSound Moment provides a further opportunity to test this thesis' design guidelines. Logs of users' music listening during product testing were made available for analysis, though were limited to weekly snapshots due to the ongoing development of the product. The music selections made in these logs were aggregated by selection style (engagement level on MoodWheel or other, such as album lookup).

RESULTS

The selection behaviours of four users participating in the testing campaign are depicted in figure 8.3. The three levels of the MoodWheel (*core*, *familiar* and *discovery*) are coloured blue and reflect mood-based selections at differing levels of engagement. *Other* methods of music selection are represented in the black column, covering the various features of the Moment, typically being album based music selection. Even with this snapshot of a few users, differences in interaction styles are clear. User 2 primarily used the MoodWheel to make *core* selections from their collection, and often made *familiar* selections to hear recommendations seeded by their collection. User 4's selections were markedly different, they used the MoodWheel at the *discovery* level for broad mood-based recommendations, and used other selection methods otherwise. Users 1 and 3 made less use of the system, with user 3 opting not to use the mood-based selection method to any significant extent.

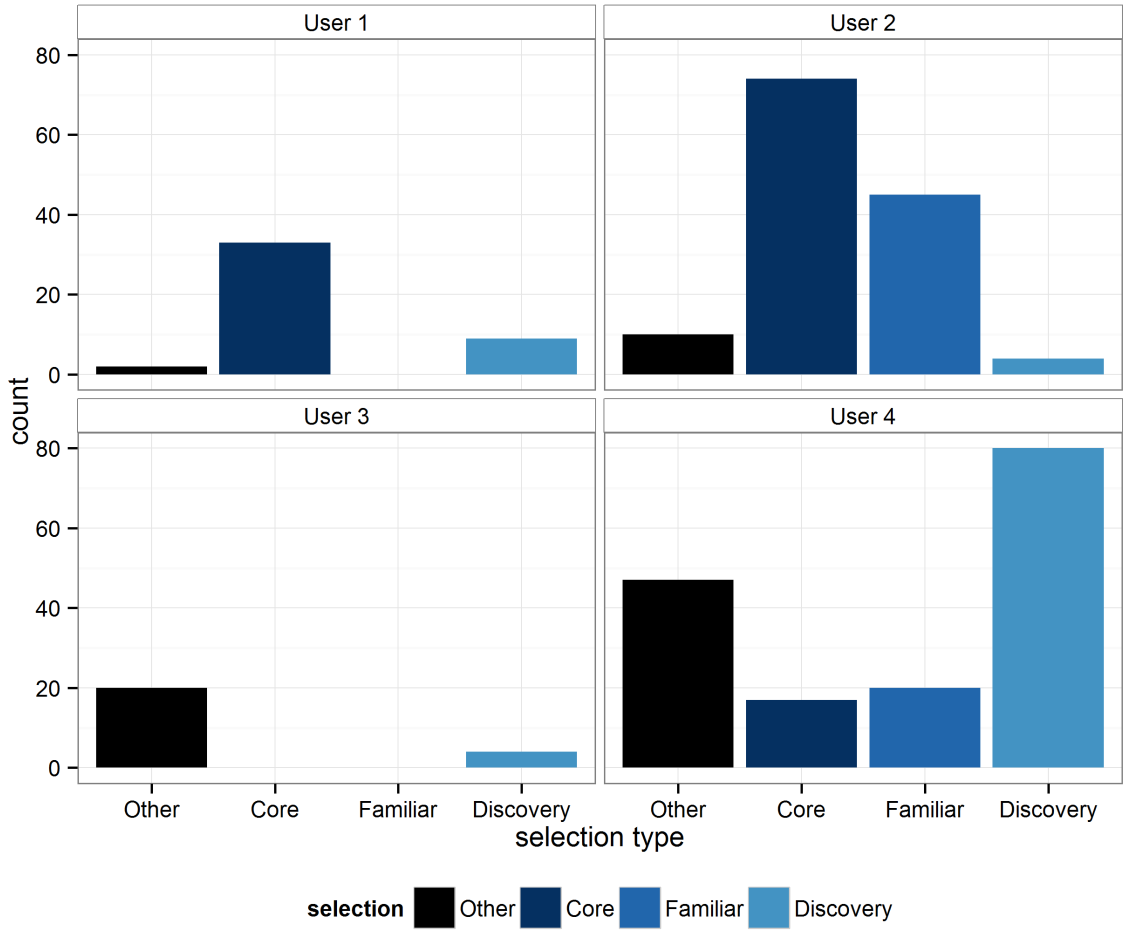

Figure 8.3: Users' choice of engagement level when using the BeoSound Moment MoodWheel over a snapshot period of several days during product development. A range of behaviours (and listening styles) are shown, from user 2's mood-based selections of mostly their own music, user 4's exploratory listening, and 3 and 4's use of other browsing methods.

8.3 EVALUATING RECOMMENDATION DIVERSITY

The information-theoretic measures developed in chapter 5 are applied here to compare the diversity of the music recommendations across the different engagement levels. The entropy over moods, tracks, albums and artists was calculated across the music selections made in each of the engagement levels. Users were able to choose the mood of music they received recommendations of, thus the diversity of mood is a reflection of the users' choices.

As this evaluation is focused on the mood-based recommendation, any track selections not made using the MoodWheel were discarded. This left 256 play events for the snapshot period available for analysis. In order to normalise the entropy (which may otherwise increase with number of track plays), relative entropy is calculated by dividing by the maximum theoretical entropy for the number of track plays for each condition. Further details on this approach are given in chapter 5. 10-fold stratified resampling was used when calculating the relative entropy, to indicate the robustness of the results despite the small sample size. The variance indicates the sensitivity of the measures to outliers and should not be interpreted as a standard confidence interval, as the resampling results can not be assumed to be normal or i.i.d.

RESULTS

The entropies for track, album and artist show the diversity of recommendations made to users. The mean relative entropies across the folds, and the standard deviations in percentage points, are given in table 8.4. These results are also depicted graphically in figure 8.5. As might be expected, the diversity of tracks from the *core* level (i.e. from the user's own music collection) was lower than the diversity of tracks from the recommendation levels. It is notable however that a more diverse range of artists was recommended at the *familiar* level than for the *discovery* level. These selection levels involve two distinct recommender strategies and so differences are to be expected, however users would likely expect to receive recommendations of a wider range of artists at the discovery level. This analysis is part of an ongoing development process, but illustrates how this approach can identify areas for investigation.

Feature	Core μ % (σ p.p.)	Familiar μ % (σ p.p.)	Discovery μ % (σ p.p.)
mood	59.3 (0.52)	53.4 (0.93)	60.94 (1.14)
track	75.47 (0.39)	94.69 (0.77)	91.69 (0.58)
album	73.75 (0.65)	93.54 (0.88)	82.06 (0.62)
artist	72.47 (0.57)	93.54 (0.88)	74.18 (0.93)

Table 8.4: Mean relative entropy (%) for each feature across the different engagement levels (Core, Familiar and Discovery) of the BeoSound Moment's MoodWheel. Standard deviation (in p.p.) is also given. These values reflect the diversity of music recommendations given for each of the recommendation strategies.

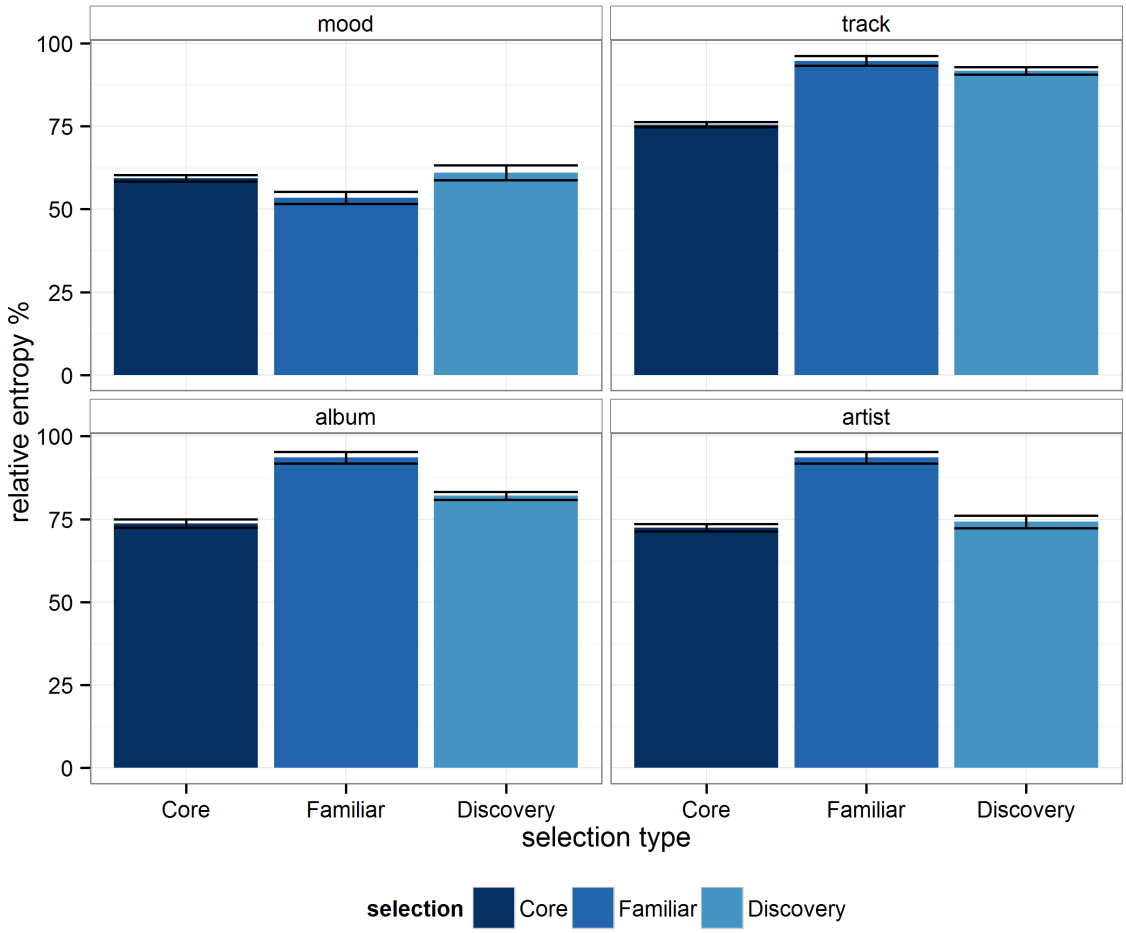

Figure 8.5: Relative feature entropy was calculated across the music selections and recommendations made in each of the engagement levels of the BeoSound Moment's MoodWheel. The bars are not standard confidence intervals but the 95% values obtained from a 10-fold stratified resampling, to indicate statistic robustness.

8.4 SURVIVAL ANALYSIS OF A MUSIC STARTUP

User engagement with a service can be measured in terms of their continued usage. One proposed approach to measuring this is through the use of Survival Analysis, as discussed in section 3.4. The work by Dupret and Lalmas (2013) and Kapoor et al. (2014) sought to capture the return time of users, linking frequent returns to high engagement. This section explores the use of Survival Analysis to indicate the 'health' of a service over time.

A dataset of logs for 1.7 million users was made available by a music retrieval startup, covering a period of 3.5 years. The survival times of each unique user was acquired from these logs, i.e. the length of time until their final use of the service. The service was discontinued at the end of the logging period, thus there is no issue of 'censoring' where some data is missing due to some users still being 'alive'. The rate at which users joined and left the service can be seen in Figure 8.6, which depicts cumulative joins and 'deaths' as well as the size of the user population over time. A key point of interest is annotated at $t = 500$ days, where the rate of users joining and leaving the service increases for a period, with the user population declining thereafter.

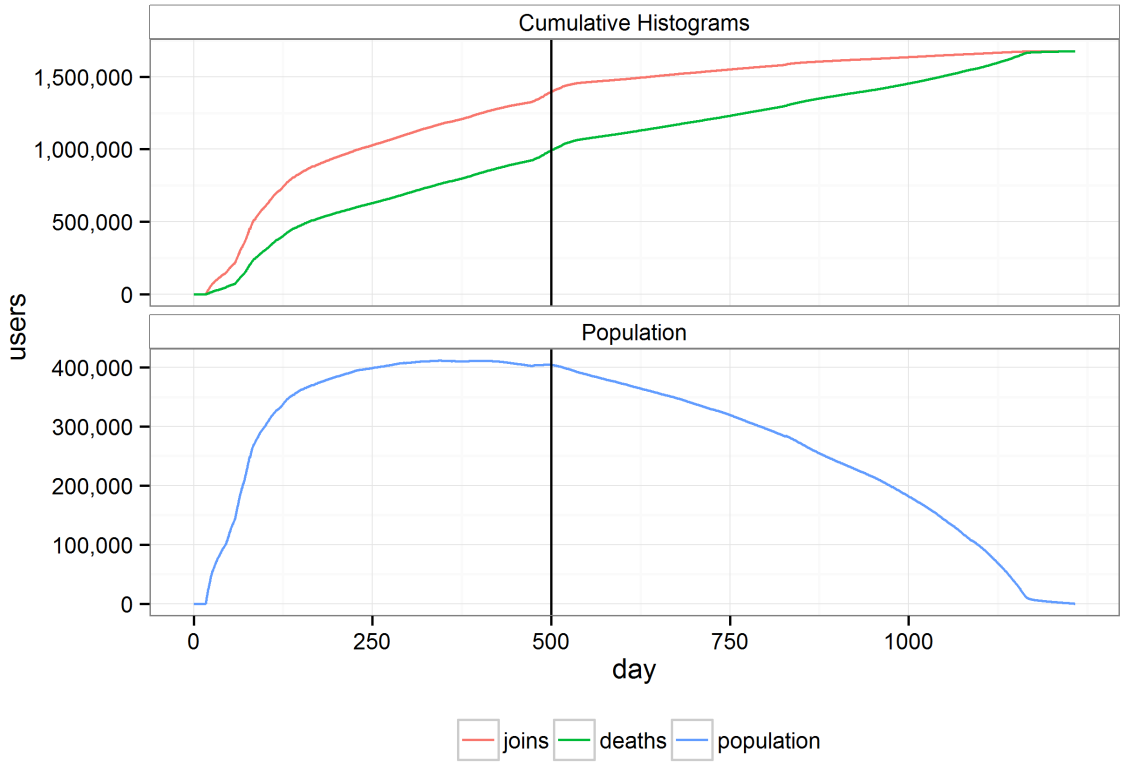

Figure 8.6: Cumulative histograms (top) depict the users joining the service and the 'deaths' as users leave the service. The population plot (bottom) shows the changes in the user population size over time.

MODELLING HAZARD AND SURVIVAL

Further detail can be gained by investigating *cohorts* of users, where users joined the service around the same time. This approach is inspired by similar analyses of animal populations, where survivorship curves capture the animals' survival probabilities over their lifetimes. Three cohorts were used to investigate the changes in how users engaged with the music service, with each cohort representing a period of 20 days around a point in the service's lifetime (100, 300, 500 days). For each cohort, 30,000 users were sampled from the period to ensure equivalent comparison.

The initial fall in survival seen in Figure 8.7 represents the *adoption rate*, with less users choosing to adopt the service in their day-to-day music listening in the 500 days cohort. The scale of the change in user engagement for the 500 day cohort is emphasised by those new users immediately falling below the number of users remaining from the earlier cohorts having joined some 200 or 400 days prior.

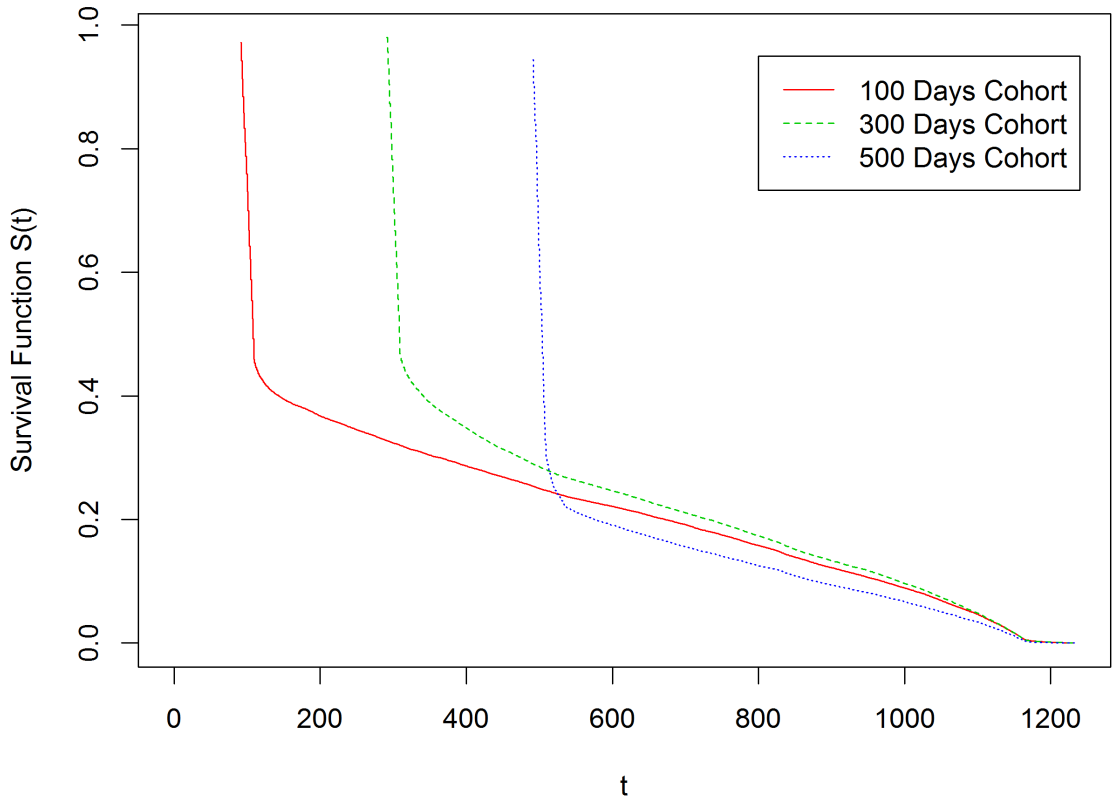

Figure 8.7: The survival function captures the proportion of users still active for each cohort. The earlier cohorts have an adoption rate >40%, which tails off slowly. The 500 days cohort has a much lower adoption rate, immediately dropping to less users than remain from the earliest cohort.

An alternative view of these cohorts' engagement is to compute the cumulative hazard function, as in Figure 8.8. This figure is logarithmically scaled, which gives a clearer view of the hazards faced by a given user (accounting for the exponential decay in users). The survival and hazard functions shown in these figures can be acquired using the Kaplan and Meier (1958) estimator, which in the case of no censoring, is simply derived from the empirical distribution function. Cox (1972) proportional hazards model regression can be used to identify factors related to the differences in survivorship, though such relationships are commercially sensitive.

The cumulative hazard of a user leaving the service is immediately higher for the users in the 500 day cohort than for the users of the previous cohorts, which has been accumulated over many hundreds of days. Potential causes for this change in new users' engagement with the service could relate to the entrance of competitors into the market. The cumulative hazard rises sharply after the announcement of the service's discontinuation, until the total cumulative hazard of every user having left.

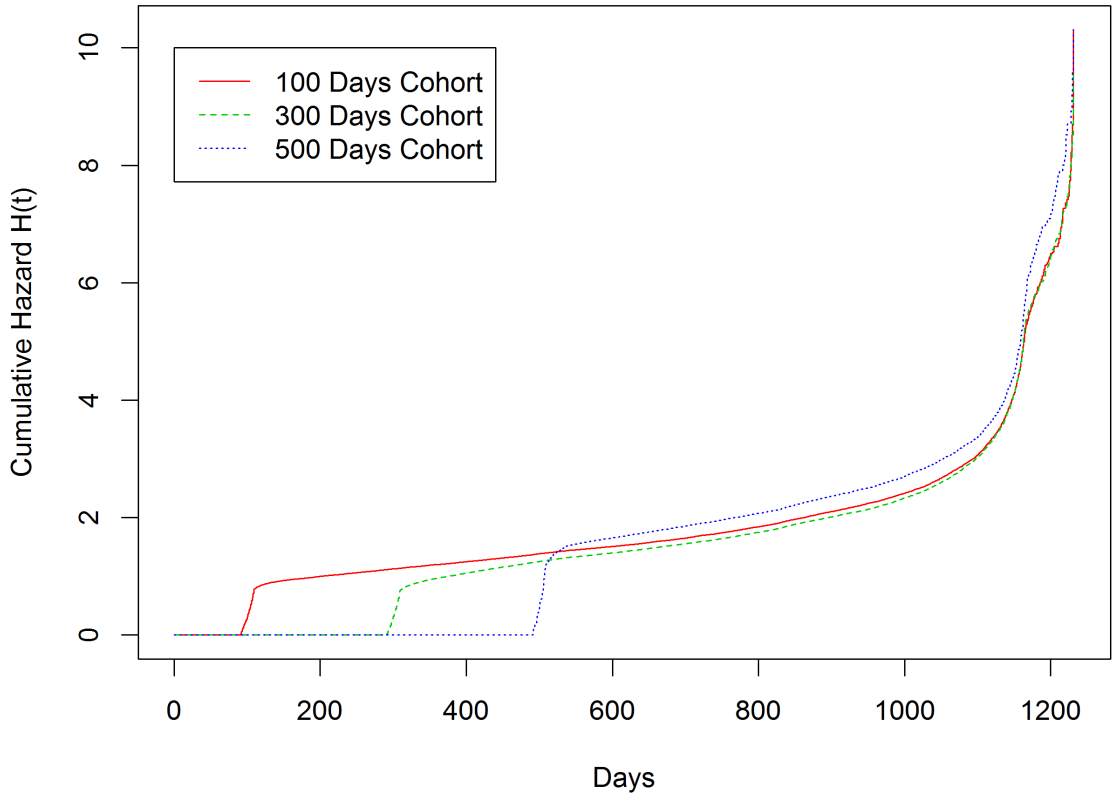

Figure 8.8: The cumulative hazard function captures the hazard or risk of leaving that each user is exposed to over their time with the music service. The 500 days cohort is immediately exposed to a greater risk of leaving than the accumulated hazards from the previous cohorts, indicating a change in new users' relationship with the service.

COMMERCIAL INSIGHTS

The key insight of the analyses of the start-up's user behaviour in this section is the point marked at day 500, where the population size began to decrease. This change is highlighted in the cohort survivorship analysis, with the striking contrast in adoption rate amongst the cohort of users joining around the 500 day point. Identifying such events and differences in user cohorts can provide valuable commercial insight. While the details of the event in this case are confidential, the effect was due to a disruptive competitor entering the market.

In the example shown, cohorts were defined in terms of the dates that users joined the service. Further insights could be gained by considering user cohorts in terms of other contextual variables, e.g., age or location. In addition to looking at how long users continued their use of the service, these techniques have also been applied to the return time of users to a service, based on the hypothesis that more engaged users will return to a service more frequently (Dupret and Lalmas, 2013; Kapoor et al., 2014).

8.5 CONCLUSIONS

This chapter has explored some of the ways in which the work of this thesis can have an impact in industry. This Ph.D. work was funded by Bang & Olufsen and thus was part of an ongoing dialogue with their designers and researchers. The idea of an engagement-dependent mood wheel interface was presented as an outcome of this thesis work, and subsequently proved influential in the development of the BeoSound Moment. The evaluative methods and measures developed in chapter 4 were used to conduct a preliminary evaluation of the implemented product, yielding initial feedback as well as providing an approach for use in continuing evaluations.

A further application of engagement-based evaluations was then explored, using a dataset provided by a music start-up featuring millions of users. In this case, user engagement was considered in terms of survival, capturing the duration of time that users remained engaged with a service. Through the use of Survival Analysis techniques, a key event was identified that had a significant effect on users' engagement with the service. This approach provides a large-scale, longitudinal view of user engagement with a music retrieval system.

9. Engagement with Immersive Media

IMMERSIVE media, where the user's attention and cognitive resources are focused on their media experience, present an example of where a more nuanced consideration of user engagement is required. The user's *interaction engagement* with the media system may be diminished by their engagement with the media itself. When users are immersed in watching a film, they may be disinclined to use an Electronic Programme Guide. Similarly, a music listener may want to play an album or playlist so they are then free to immerse themselves in the music. A rather extreme example of this tension is the case of Virtual Reality (VR) Head Mounted Displays (HMDs) – while receiving a great deal of industry investment, the current interaction scenarios greatly impede the user's ability to interact with the system via established modalities such as keyboard, mouse, etc. VR provides a context for considering more generally the competition for the user's attention, with engagement not simply the level of effort from the user, but how they apportion that attention and effort.

9.1 BACKGROUND

When engaging in an interaction with a VR system, the user's control has predominantly been through the use of gestural interfaces (such as the Leap Motion) or interfaces that are suitable for use without sight (tangibles, handheld controllers, or relying on a subset of keyboard/mouse commands) (Billinghurst and Kato, 1999). These solutions, while potentially adequate for low levels of *interaction engagement*, are almost certainly inadequate for higher levels. Tasks such as text entry, drawing on a pen display etc. require a greater bandwidth of input and a high *interaction engagement*, making them difficult to perform using a HMD. In particular, interactions break down where the output feedback is not incorporated into the VR presentation, e.g. trying to use a phone while wearing a HMD.

Keyboards are a ubiquitous input technique and a key example of the issue of limited *interaction engagement* with real-world peripherals when using VR HMDs. Their use involves both kinaesthetic feedback from fingers striking keys and visual feedback to observe hand and key placement. Virtual keyboards offer a replacement input modality in virtuality, but lack haptic feedback. This can have a dramatic effect on typing performance. In a study of typing performance on a flat keyboard (lacking the haptic feedback of key travel) versus a standard keyboard in VR, Barrett and Krueger (1994) demonstrated a significant performance drop without haptic feedback. A variety of text entry methods for use when wearing VR HMD (both mobile and static) have been proposed. Outside of speech, none have been shown as approaching the performance of the standard PC keyboard (González et al., 2009). There are many contexts in which speech may not be appropriate, not only for social reasons but also the twitch responses required in gaming. This chapter thus investigates how the user's engagement with their real-world peripherals such as keyboards could be supported when they are engaged in immersive VR media experiences.

Visual feedback has also been shown to impact typing rate and accuracy. Clawson et al. (2005) evaluated keyboards for blind-typing, using a small mini-QWERTY keyboard. They found that when users had no view of their hands typing, their typing rate and accuracy suffered. They note that when deprived of on-screen visual feedback, users' typing rate actually increased. They term this the 'Skywalker effect', i.e. on-screen feedback can impair performance by making users over-conscious of their errors. This result is also interesting in that the user attending to the feedback, having a high *interaction engagement*, actually impeded task performance – again showing the need for models of engagement rather than simple metrics of performance. In earlier work, Rosinski et al. (1980) showed that typing rate itself was not much influenced by the on-screen feedback, but the number of corrections was. These studies confirm the model set out by Long (1976) – haptic and visual feedback support typing while on-screen feedback is a higher level control loop, only supporting correction.

For high bandwidth input devices like keyboards, performance suffers without the low-level feedback present in reality (haptics and visual). The on-screen feedback in the studies did not support the high-bandwidth, low-level input. When using a virtual keyboard in VR, users have no 'real' feedback, with only 'on-screen' feedback available. It is clear that it is desirable to provide such feedback to the user, however doing so affects the user's immersion in VR. This issue is addressed using an engagement dependent approach in section 9.3.

IMMERSION AND PRESENCE

Immersion in VR is typically quantified through the user's sense of presence. Presence can be affected by the rendering quality of the scene, the quality of the HMD's head tracking, and even how users interact with virtual objects. It can be increased through natural interactions with objects as they occur in the real world (Kaber and Zhang, 2011). There are a number of approaches to measuring presence, such as instrumenting the user to monitor brain activity or more traditional qualitative measures (Van Baren and IJsselsteijn, 2004). The Igroup Presence Questionnaire (IPQ) (Schubert et al., 2001) is widely used and sufficiently generalised to be adopted in this chapter. Slater (2009) described presence as *place and plausibility illusion* – by taking an engagement-dependent approach to incorporating feedback from reality into VR, the impact of such blending on place illusion can be minimised.

MIXED REALITY

The field of mixed reality considers displays as being on a continuum, spanning from reality to virtual reality – the *Virtuality Continuum* (Milgram and Kishino, 1994; Milgram and Colquhoun, 1999). This continuum led to the definition of both Augmented Virtuality (AV), where a virtual reality or *virtuality* view is augmented with elements of reality, and Augmented Reality (AR), where a reality view is augmented with elements of virtuality. A display's position on the continuum is determined by the balance of reality or virtuality that is incorporated into the display feedback. AR has seen increased research focus in recent years, building upon early work such as Navicam (Rekimoto and Nagao, 1995), with reality being augmented to allow for novel interactions. AV has also been investigated, through the use of chroma-keying approaches blending elements of reality into a VR space. Head-mounted cameras have often been used to capture reality, e.g. Steinicke et al. (2009)'s use of chroma-key techniques to incorporate the user's body, presenting a virtual hands and body in an egocentric view of virtuality. More recently, head-mounted depth cameras have allowed for incorporating hands or objects into virtuality, such as in Tecchia et al. (2014). The availability of larger area sensing, such as with Microsoft's Kinect, allows for multi-user tracking, as well as gestural interaction, which could be used to more broadly augment the virtuality environment.

There remains the question of how best to manage the augmentation of virtuality - managing traversals or transitions within the virtuality continuum as discussed in Davis et al. (2003). Mixed reality research such as Benford et al. (1998) has explored boundaries, where physical boundaries mark the change between mixed and virtual reality. Where an interaction spans distinct points on the mixed reality continuum, it is termed a *transitional interface* (Grasset et al., 2005). The importance of continuity in transitional interfaces has been highlighted, as well as the lack of a theoretical guideline for when to make these transitions, and how much to transition by (Carvalho et al., 2012).

9.2 LINKING ENGAGEMENT TO MIXED REALITY

Chapter 3 explored a number of conceptions of user engagement, and this thesis has developed engagement-dependent approaches to retrieval and interaction, such as in chapter 7. From the control-theoretic perspective, the more engaged with an interaction users are, the more frequently and accurately they sample feedback and the more they provide rapid and accurate high-bandwidth input. While the literature review and findings of chapter 4 highlight the need for a more nuanced understanding of engagement, extending to affective and other factors, the control-theoretic model does allow for a coherent interaction to be designed that can span levels of engagement.

Depending on their current context in VR, users will vary in how much they wish to engage with interactive objects in reality. Users may wish to engage with and incorporate real persons around them or interactive devices like keyboards or phones into their VR experiences, e.g. Bruder et al. (2010) used chroma-keying to incorporate real-world tools into VR according to architectural context when exploring a VR house. Transitions in mixed reality can be linked to user engagement with interactive elements in the different environments. A user exploring a VR environment may have no need for objects in reality. However, if they extend their hands to engage with their keyboard, there should be a transition to a mixed reality mode. Allowing users to control their position on the mixed reality continuum combines transitional interfaces, mixed reality and the engagement-dependent approach.

While this engagement-dependent blending allows users to see and interact with real-world objects only as they wish to engage with them, there remains the question of identifying this level of engagement. Engagement was denoted explicitly by users in chapter 7. This chapter explores implicit measures of engagement such as gaze, as well as more explicit measures such as the user's gestures to interact with the device. It is also likely that users would wish elements of reality to be blended when there is a *likelihood* of engagement, for example a user with a VR HMD could be made aware of others entering the room. The issue of personal space and the social contexts around VR relate to the affective aspects of engagement not captured in the control-theoretic view.

This chapter aims to show how an engagement-dependent approach to AV supports richer interactions with higher bandwidth input. Figure 9.1 shows the control loops involved in the presented engagement-dependent, mixed reality system (blue for real, red for VR). The VR control loop involves input from reality and feedback from the virtual environment. The VR control loop is high-level, comparable with on-screen feedback for typing: users see the results of their typing but not their hands. The more a user engages with a real interactive object such as a person or keyboard, the more it can be blended into the augmented virtuality view. Users can thus perceive the low-level feedback from interacting with the real object, e.g. hitting keys when typing or social signals from proximate persons. They are able to choose to engage with reality from a virtuality context, with their control loop with reality becoming more tightly-coupled so that they can rapidly sample detailed feedback from reality.

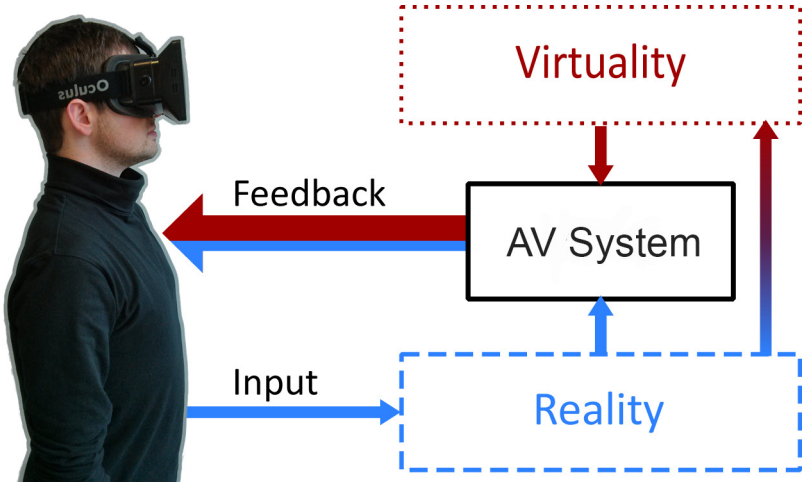

Figure 9.1: Feedback from real objects being engaged with can be used to augment virtuality, providing a low-level control loop (in blue) that can support high bandwidth interaction, such as typing. As the user engages with reality, the more real feedback is mixed with virtuality.

9.3 MIXED REALITY TYPING STUDY

To interact with a VR environment, a user requires an appropriate input modality. Input actions occur in reality and in VR the user loses the inherent visual feedback for these actions. Many text entry methods for VR have been considered (González et al., 2009), with voice input offering reasonable performance, however adoption of novel modalities will be slow and the use of keyboards is likely to continue, even if only transitionally. The keyboard is a ubiquitous input device and widely used by gamers (currently the key demographic for consumer VR HMDs); it is familiar, its layout is consistent, it has guidance for blind typing, and users are proficient in its usage. The use of a keyboard in VR presents the immediate problem of requiring the user to locate it in reality while immersed in VR. Users must switch from considering their virtual environment to the spatial layout of their real surroundings.

Typing offers an interesting case for examining the ability to interact with reality, being an example of a rich interaction with an object that requires a high-bandwidth feedback loop. Enabling high performance keyboard use is a quantifiable means of demonstrating the general opportunity for rich interaction supported by engagement-dependent AV. This chapter explores bringing the real keyboard into virtuality, hypothesising that this would demonstrate increased performance when blending feedback from reality with virtuality.

DESIGN

A text entry study was conducted with four conditions. It was designed to assess both the status quo performance with no view of reality, as well as the potential for improving performance through incorporating either a full view of reality, or a selected subset of it.

1 Reality Baseline typing performance on a keyboard with full view of reality (no HMD).

2 Virtuality Status quo with no keyboard view and wearing a HMD (typical of current VR HMD users' setups), as in Figure 9.3.

3 Partial Blending A view of the keyboard and user's hands was blended into virtuality based on engagement, e.g. reaching out for the keyboard, as in Figure 9.4.

4 Switching The view was switched between reality or virtuality based on engagement, as in Figure 9.5.

Figure 9.2: Experimental setup for the VR typing study. Users wore a Oculus Rift VR HMD augmented with a camera for mixed reality blending. Users entered text using a standard desktop QWERTY keyboard. The desk and area in front of the user was made green to allow a chroma-keying approach to the mixed reality blending.

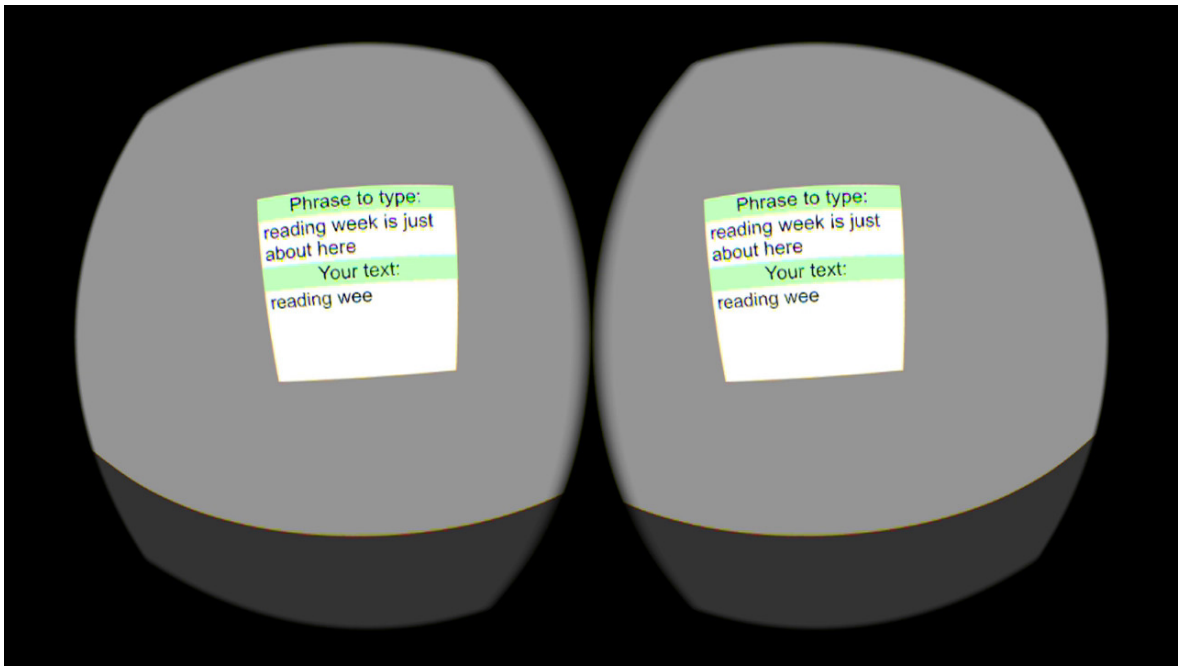

Figure 9.3: Typing Condition 2: No mixed reality blending. Users were unable to see the keyboard and instead could only see the VR environment, in this case the text entry prompt.

Figure 9.4: Typing Condition 3: Partial mixed reality blending. Users were immersed in the VR environment however elements of reality were blended in as engaged with, e.g. as the user reached for the keyboard.

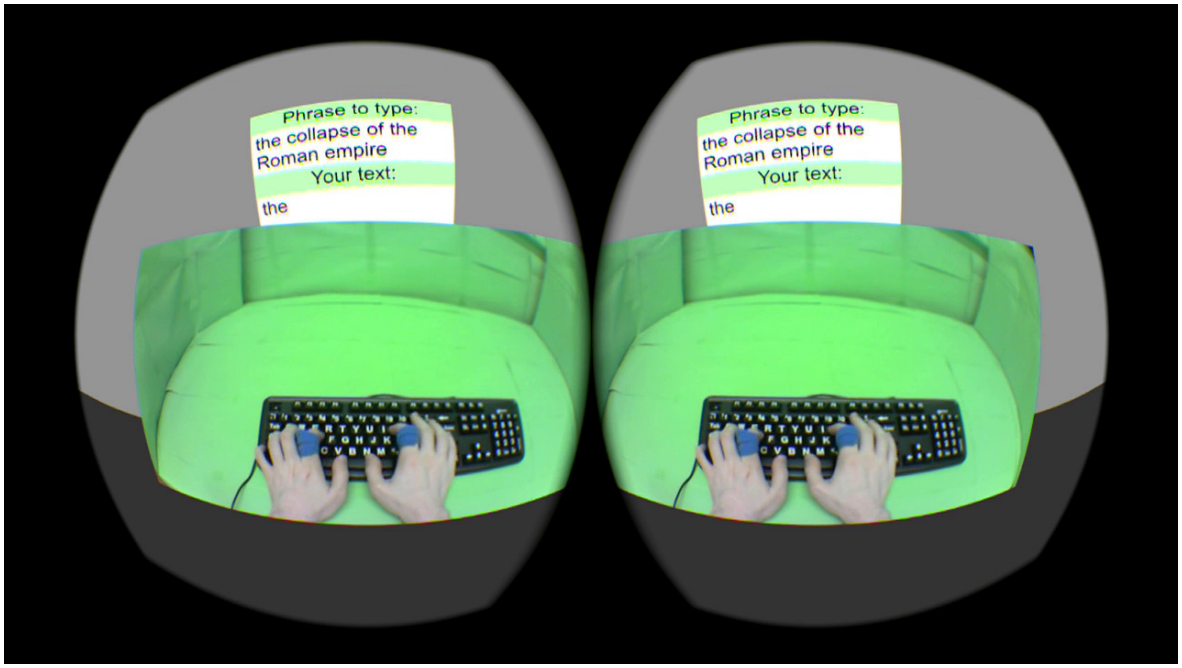

Figure 9.5: Typing Condition 4: VR/Reality switching. The user's view switched between reality and virtuality depending on which they were presently engaging with. Note that although the camera view doesn't appear to fill the display, it does fill the field of view.

Sixteen participants were recruited from University mailing lists (Age mean=25.6, SD=4.0, 14 male, 2 female) in a within-subjects design. For the baseline condition, users were seated at a standard single monitor workspace, while in virtuality users found themselves in a VR space with the same view as was on the monitor previously. Users were asked to enter 15 phrases, with 1 training phrase at the start of each condition to familiarise them with the keyboard view presented. Before each phrase, participants were asked to put their hands by their side, in order to mimic aperiodic interaction with the peripheral from VR. Based on Kristensson and Vertanen (2012), the MacKenzie 500 phrase set was used, with phrases chosen at random and presented at the top of the display, with the bottom of the display used for on-screen feedback of typing. Baseline typing was captured first, the remaining conditions were counterbalanced. In addition to standard typing metrics (Soukoreff and MacKenzie, 2003), accuracy and time to first key press were also measured to gauge how effectively participants could locate the keyboard.

IMPLEMENTATION

Bringing the keyboard into VR was achieved with the use of a modified Logitech C310 HD webcam, with a 1.8mm M12 wide-angle board lens, mounted on the front of an Oculus Rift DK1. A single camera setup with only monocular depth cues was used as this approximates current camera setups in VR headset/phone hybrids (such as the Samsung Gear VR), as well as minimizing processor load and latency. The scene was rendered at 60FPS.

For selectively blending reality, a chroma-key approach (Figure 9.2) was utilised whereby the keyboard was placed within a green screen environment, with multi-threaded image processing and HSV thresholding in openCV used to determine the green screen contour and non-green contours within it, which were then presented in virtuality. This was performed separate to the rendering of the webcam, to minimise the latency of the user's view of reality, through an alpha mask which was generated and applied with ~1 frame of latency. Hand detection occurred within the same process, with users wearing markers on their hands that were detected via openCV blob detection. This was implemented in the Unity 4 engine. The VR scene was rendered on a Intel Core i5 (3.30GHz) with an Nvidia GTX 460 GPU. The keyboard used was a standard PC size and layout, however it featured larger lettering; this was to compensate for the low resolution of the Oculus Rift HMD.

RESULTS

The typing statistics can be seen in table 9.6 and include a repeated-measures ANOVA (GLM) with post hoc pairwise Tukey's tests ($\# - \#$ indicates a significant difference between numbered conditions, $p < 0.05$). Performance was highest in the '1 Reality Baseline' condition for most metrics, in keeping with previous work in the literature. The status quo condition of '2 VR No Blending' had significantly higher error rates (corrected and not corrected) and a large drop in typing rate (WPM). The two AV conditions (3 and 4) significantly reduced the error rate in VR typing, nearing the baseline rate of typing in reality. Typing rate however, while improved, failed to reach reality baseline performance. There was no significant difference between partial and full blending in these measures. Figure 9.7 demonstrates the behaviour participants employed in order to type as quickly and accurately as possible: for condition (2) where users had no view, they quickly made inaccurate key-presses for orientation to the keyboard, whereas the AV conditions (3 and 4) introduced some delay.

DISCUSSION

Using the engagement-dependent approach to selectively augment virtuality with the real keyboard greatly reduced error rates. Users were able to orient themselves to the keyboard as if it were in VR, they did not need to attempt to remember where the keyboard was in the real environment while in VR. Typing rate is a low-level feedback loop and not supported by the usual feedback in VR. Bringing the visual feedback of the real keyboard into VR helps, however a lack of stereoscopic depth cues and the latency of the VR HMD likely contributed to the modest gains in WPM in the AV conditions. These results demonstrate that incorporating reality into VR is necessary to preserve performance, and that only providing a view of the keyboard and hands does not negatively impact performance versus a full view of reality. The engagement-dependent approach of transitioning the control loop between AR and VR enables this partial blending of reality, supporting rich interactions in mixed reality.

	1 Reality Baseline	2 VR No Blending	3 VR Partial Blending	4 VR Full Blending	RM ANOVA	Tukey post-hoc
WPM	58.93 (17.03)	23.6 (22.34)	38.47 (19.34)	36.63 (19.74)	F(3) = 30.55, p <0.01	2-1, 3-1, 4-1, 3-2, 4-2
Duration to First Key (sec)	1.92 (0.53)	3.12 (1.89)	3.27 (0.79)	3.42 (0.86)	F(3) = 9.07, p <0.01	2-1, 3-1, 4-1
Total Error Rate	4.64 (3.09)	30.86 (15.17)	9.2 (6.43)	10.41 (7.36)	F(3) = 28.45, p <0.01	2-1, 3-2, 4-2
Not Corrected Error Rate	0.87 (1.55)	0.61 (1.37)	1.07 (1.41)	1.78 (3.52)	F(3) = 1.36, p = 0.27	
Corrected Error Rate	3.77 (2.51)	30.25 (15.37)	8.13 (6.03)	8.63 (7.37)	F(3) = 73.32, p <0.01	2-1, 3-2, 4-2
First Key Correct	97.77 (4.3)	49.11 (24.45)	89.73 (16.28)	91.96 (14.49)	F(3) = 29, p <0.01	2-1, 3-2, 4-2

Table 9.6: VR typing results. Total Error Rate from Soukoreff and MacKenzie (2003). First Key Accuracy is between 1 (100% accurate) and 0 (0% accurate). Tukey's tests show statistically significant pairwise differences between conditions, $p < 0.05$.

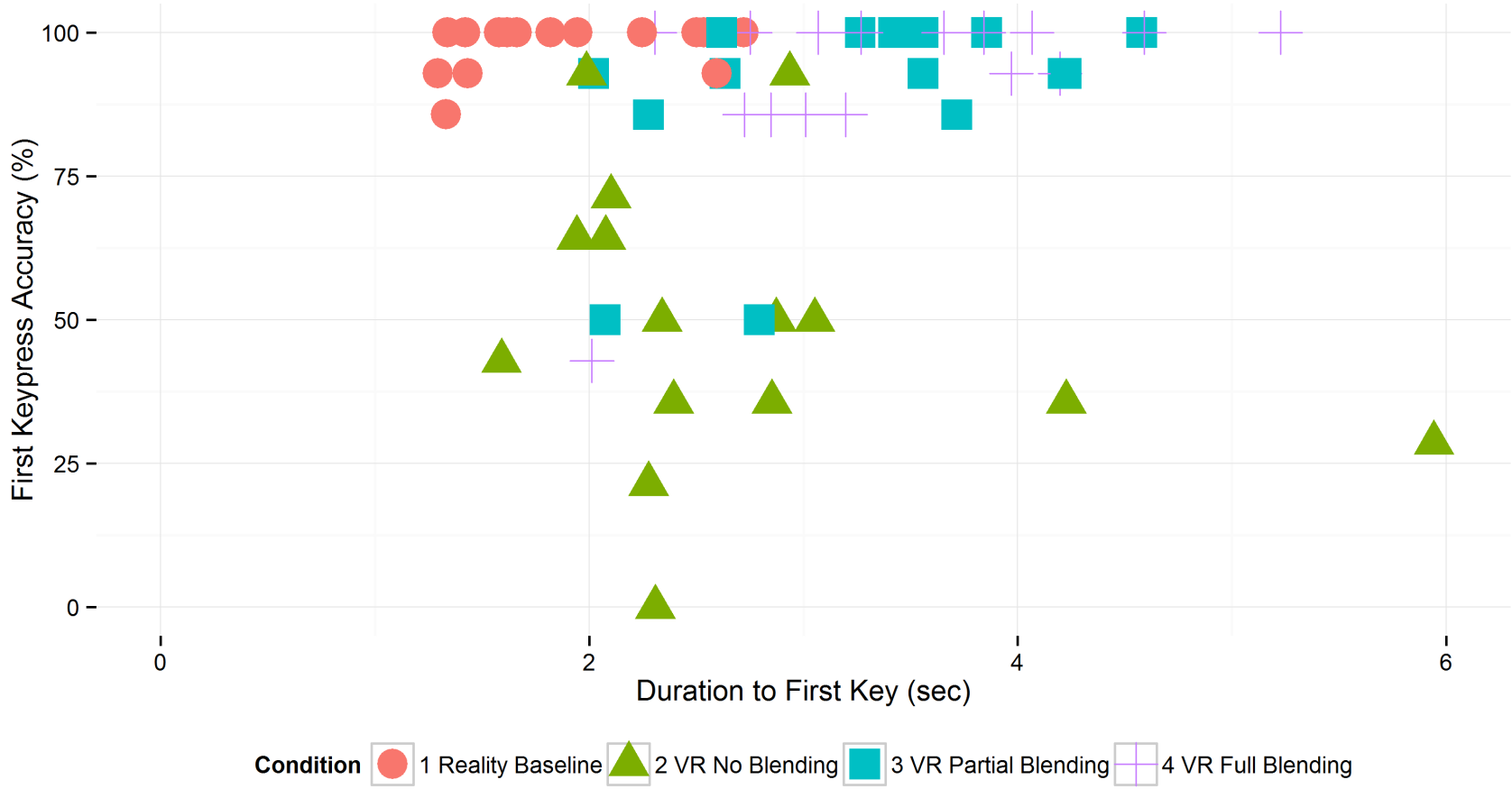

Figure 9.7: VR first keypress accuracy and delay across conditions. Users in the '1 Reality Baseline' condition had no view of the keyboard, they quickly made inaccurate keypresses to orient themselves. This contrasts with users in the AV (3 and 4) conditions, largely achieving high accuracy but with some delay introduced.

9.4 ENGAGEMENT-DEPENDENT MIXED REALITY

The aim of this chapter was to exploit engagement-dependent approaches to enable a broad range of interactions with both reality and virtuality, as the user engages with elements of either. The typing study provided a quantitative insight in how these mechanisms impact the performance of interaction. While typing is useful for media retrieval, in particular when producing textual queries, it remains somewhat tangential to the larger goal of considering engagement with VR media. This section briefly discusses the use of engagement-dependent AV to enable a range of mixed reality interactions as part of a media experience. The impact of these approaches on the user's sense of presence and immersion in their VR experience is explored in further studies by McGill et al. (2015). An online overview video of the work in this chapter, as well as the additional studies, is available[1].

INCORPORATING OBJECTS

The VR desktop setup depicted earlier in Figure 9.2, and used for incorporating a keyboard into VR, can also be used to incorporate other desktop objects. Figure 9.8 depicts the blending of a number of desktop items related to media interactions: the keyboard for querying and control, as well as food and drink. Three engagement-dependent AV strategies were explored. *Minimal blending* incorporates elements around the users' hands as they bring their hands into view, allowing the user to explore the real space and use objects while minimising impact on immersion in VR. *Partial blending* incorporates all interactive objects likely to be engaged with as the user extends their hands to interact. *Full blending* reflects the status quo in industry where gestures are used to switch between a real or a virtual view.

INCORPORATING PEOPLE

In a survey of VR users, McGill et al. (2015) identified that the user's lack of awareness of people around them is a pressing issue. An interesting aspect of this issue is that the user's engagement with others is a social, affective aspect of engagement and not the easily quantifiable interactions thus far considered in this thesis. Another issue is that a user is unable to choose to engage with a person if they are not first made aware of their presence. There is thus an implicit baseline of engagement with others, who are represented in Figure 9.9 as ghost-like figures. As these figures are engaged with, detected in terms of gaze or conversation, the alpha value on the blending is changed such that the ghost-like figures become fully present in the VR environment.

[1] `http://www.dannyboland.com/CHI-VR/` (20/02/15)

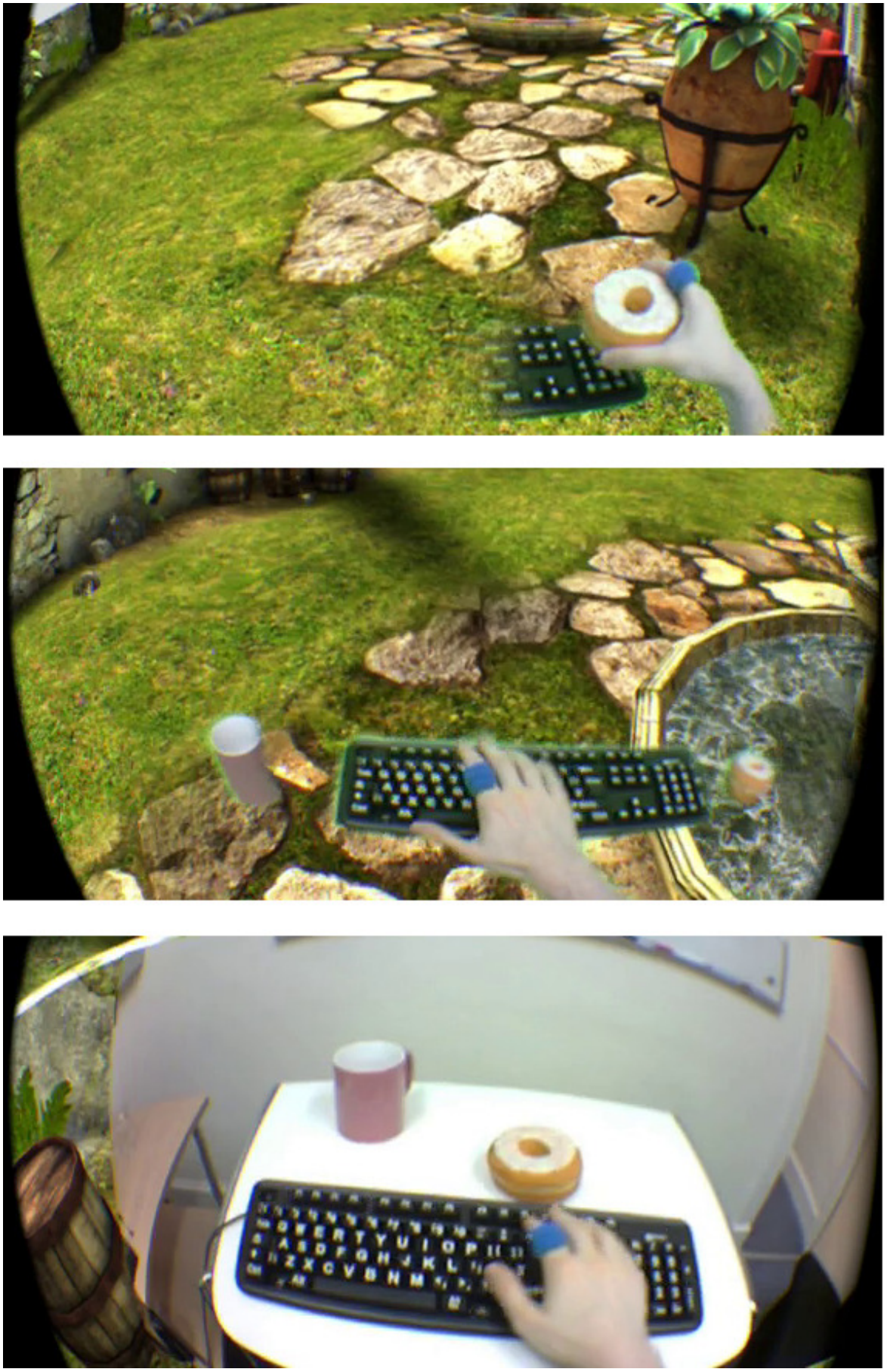

Figure 9.8: Top: *Minimal blending* (reality around user's hands). Middle: *Partial blending* (all interactive objects). Bottom *Full blending* (all of reality).

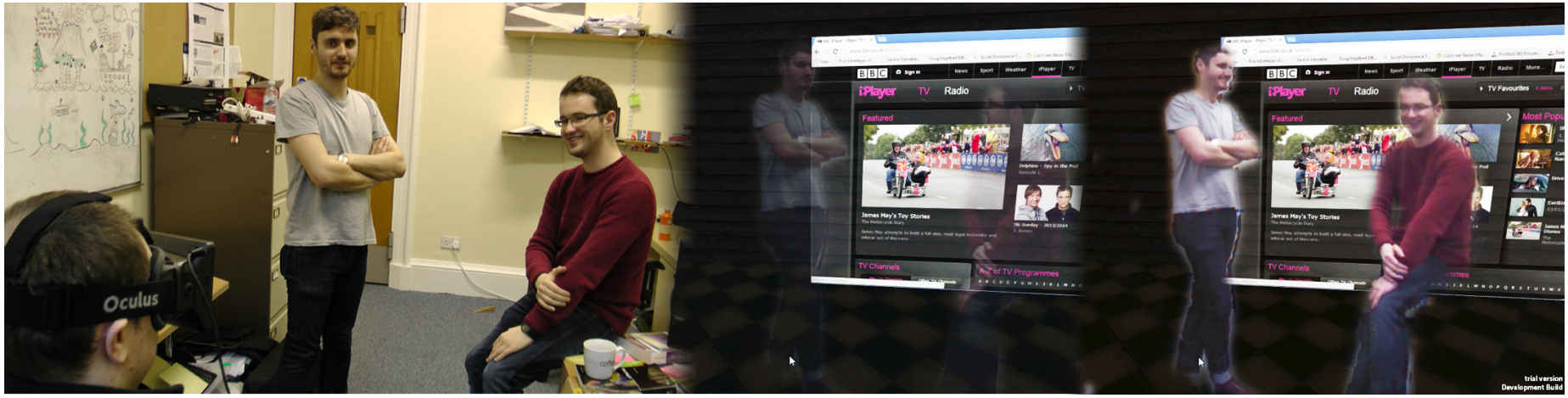

Figure 9.9: People in the surrounding real environment are incorporated into VR based on engagement. Given the implicit social engagement of a person being in the same real environment, ghost-like figures are blended into VR (middle). As the user engages with these figures, either by gaze or conversation, the alpha value of the blending is adjusted to bring them fully into the VR environment (right).

9.5 GENERALISABILITY

At the heart of the work in this chapter is an interaction with an autonomous agent, which utilises inference about user intent, conditioned upon their engagement. While this work is a departure from information retrieval, it demonstrates the applicability of engagement-dependent interaction beyond retrieval interfaces. Whether adaptivity is applied in a recommender system or an interaction with conditional dynamics, a key outcome from this thesis is the proposal of user engagement as a contextual variable for such inferential systems. Interactions will increasingly involve a negotiation with autonomous agents, whether to manage large libraries of content, a VR experience, or even an autonomous vehicle for example. There remains a need to explore the mechanisms by which users can engage or disengage in controlling such interactions, or delegate such control to an agent. This thesis was motivated by the burden of too much control - *too-much-choice*, however it will be equally important to avoid the frustration of users being disempowered by too little control.

Considering the concept of engagement-dependent interaction in the context of VR has facilitated the solution of usability issues facing VR HMD users. Users were able to interact with real objects and peripherals from the context of a VR environment, through selective engagement-dependent Augmented Virtuality. This approach enabled interaction with reality, while preserving presence and immersion in VR, by selectively blending relevant parts of reality with virtuality. This was done in an engagement-dependent manner, inferring when and how much to blend reality into virtuality based on the user's engagement with objects in reality. This approach is distinct from past work in VR, seeking to preserve presence and place illusion with a mechanism that minimises the blending of reality.

10. CONCLUSIONS

THIS thesis has motivated the need to adapt music retrieval interfaces to users' level of engagement in the retrieval process. It defines a number of quantitative measures of music-listening behaviour, which were then experimentally linked to user responses in music engagement questionnaires. Using these measures, and a synthesis of the literature on user engagement in retrieval, a number of music-listening profiles have been derived as design guidelines for the development of music retrieval systems. Two examples of such music retrieval systems were designed and developed, targeting different levels of user engagement, and evaluated in terms of how they support the listener profiles. This thesis' engagement-dependent approach to music retrieval has influenced a commercial product, which is presented along with an exploratory evaluation of how this system performs. Finally, with a view towards generalising the engagement-dependent work, the design methodology of adapting control to user engagement was applied to a novel media interaction – virtual reality head mounted displays.

10.1 METRICS OF LISTENER BEHAVIOUR

The first contribution of this thesis, which motivates a large proportion of the subsequent work, is the proposal of a series of quantitative metrics developed in part II to capture aspects of users' music listening behaviour. These metrics were grounded with a questionnaire, adapted from music psychology, which was posted on an online community for Last.fm users and completed by 91 music listeners who also provided their Last.fm listening histories. A larger analysis of how users organise their playlists was also performed, after crawling 19,225 hand-made playlists from 7,666 Last.fm users.

The first set of metrics capture relatively simple music retrieval interaction behaviours, such as album-based listening and track skipping, and were derived from first principles in section 4.2. A more subtle approach to capturing listener behaviour was then developed, where the users' speed in selecting new tracks or skipping tracks is measured, i.e. how far into a song they wait before skipping. These metrics were then correlated with the users' level of control over music selection, as indicated in their questionnaire responses. Linking listener behaviours to their stated engagement levels in this way provided much of the detail used to build the engagement-stratified profiles of music listening behaviour.

ENTROPY

Given the many features that can be derived to describe music (such as genre, mood, artist etc.), it was desirable to develop a general approach to evaluating changes in music-listening behaviour, in terms of arbitrary features. Mood was chosen as an example case, with mood features provided by Syntonetic, a commercial partner. By computing the entropy of each of these music features throughout a user's listening history, a portrait of their music listening behaviour can be constructed, as in section 5.1. Given that listening behaviour is likely to change over time, this entropy calculation was windowed over a number of tracks, allowing change-point detection to identify changes in listener behaviour.

This approach was scaled up in section 5.2 to consider thousands of users and tens of thousands of hand-made playlists, identifying which music features reflected the way in which users organise their playlists. By computing the mutual information between playlist membership and the music features of tracks, the music features that shared the most information with the playlist organisation could be ranked. While the music features used in this thesis are proprietary, they allowed an illustrative analysis to be performed at scale, and demonstrated the value of combining metadata features from a different source (Spotify) with a few of the audio features, rather than considering the full high-dimensional space of all the audio features, which demonstrated diminishing returns with each feature.

10.2 DESIGN GUIDELINES

The review of work on music listening behaviour and user engagement, in particular with music, motivated the identification of a number of engagement-stratified music listener profiles – detailing music listeners' behaviour as it relates to their engagement. These listener profiles could then be used as design guidelines in the creation of music retrieval interfaces. With the survey and quantitative measures in part II, the profiles were elucidated as follows:

Engaged These users have a high initial engagement in the interaction, with more specific retrieval queries, e.g. selecting a particular album. They then make very few further interventions, which are quick and decisive.

Casual These users wish to satisfice, investing little effort in the retrieval at any given point. Their lack of initial control means that these users need to be able to easily make corrective interventions.

Mixed Most users vary between levels of *Casual* to *Engaged* music listening, depending on their listening context.

Framing music retrieval interactions in terms of engagement provides a spectrum on which to consider existing interfaces, which typically cater to a specific level of engagement at one of the extremes. Shuffle-based playback is a common mechanic to cater to *Casual* listening, and the use of text retrieval and hierarchical menus is suited to *Engaged* retrieval. Recommender systems, and in particular those which adapt to user preference and context, provide a means to span the gulf between these extremes.

ENGAGEMENT AND CONTROL

The review of literature on user engagement in chapter 3 described distinct factors: *retrieval control* as the user's desire and effort to control the retrieval outcome, *media engagement* as the user's attention and involvement in the media, and *interaction engagement* as the user's attention and involvement with the interface itself. While users were described in the profiles as casual to engaged in terms of their *music engagement*, the measures of music listening behaviour identify the extra nuance of these factors. *Engaged* users like to have a high degree of *retrieval control* over music, as discussed in the profile and the literature, but this often takes the form of album selection, with the user then having less *interaction engagement* over the course of the interaction. Relatedly, *Casual* satisficing of songs with little *retrieval control*, for example in shuffle selection, may still result in listening with a high degree of *media engagement* or *interaction engagement*, with the listener exploring new music and making decisions whether to skip unsuitable tracks.

10.3 DESIGNING FOR ENGAGEMENT

With the music listening profiles, as well as the review of music retrieval interaction, it is clear that user engagement in their music retrieval is a key contextual variable for the design of music retrieval systems. This thesis set out to illustrate how retrieval can be designed for, and adapted to, user engagement. A series of demonstrator systems were developed in part III, to investigate and overcome technical challenges in adapting retrieval to engagement. One of these systems informed the development of a product produced by Bang & Olufsen, which provided an opportunity to evaluate a 'real-world' implementation in chapter 8.

CASUAL MOBILE RETRIEVAL BY TAPPING

The first demonstrator system (chapter 6) sought to overcome the issues of subjectivity that complicate music retrieval by allowing users to query using a fundamental aspect of music – its rhythm. A key challenge was identified in that user querying by rhythm involved the subjective sampling of instruments in polyphonic music, as well as differences in the tempo and detail of how users reproduced rhythm. A simple generative model of rhythmic queries showed how this issue could be addressed, though is dependent on polyphonic onset data. The resulting interaction afforded casual music retrieval, with users not having to remove their mobile device from their pocket to shuffle by tempo. It also allowed for retrieval interactions with more *interaction engagement* and *retrieval control*, with users more carefully reproducing the rhythm of a song for specific item retrieval.

SPANNING ENGAGEMENT LEVELS

A tablet-based system (chapter 7) was developed to demonstrate how a retrieval interaction, with a coherent interaction model, could span the range of listener profiles and engagement levels discussed. A pressure-based metaphor for allowing users to exert their engagement with the retrieval was employed; detecting engagement implicitly via sensing remains a compelling research challenge. This interface presented the user with a simple overview of their music, organised by mood, which users could swipe to steer music recommendations. As users applied pressure, engaging with the system, the system's generative model of user input assumes more precise input given the engagement context, and infers more specific recommendations. This mechanic allows the interaction to span from broad mood recommendations through to specific item retrieval, with just pressure and touch modalities. This engagement-dependent style of music retrieval was incorporated into the design of the BeoSound Moment, which is detailed with an exploratory evaluation in chapter 8.

BEYOND MUSIC

In more exploratory work, adapting a Virtual Reality interaction to user engagement was considered in chapter 9. Media interactions in VR are an extreme case of immersive media experiences, where the user's *media engagement* interferes with their ability to engage with and control the interface. The need to balance how much of the burden of control to place on the user applies here too, in order to preserve immersion in the media. The examples thus far consider only that users may engage less with a retrieval, perhaps due to distractors or temporarily limited cognitive resources, however with VR the challenge is to successfully balance engagement in the virtual environment with engagement with the modalities required to control it. The engagement-dependent interaction dynamic was applied here too, for example selectively blending a keyboard into VR as a user reaches out to interact with it. A series of studies showed that users could provide controlling input more easily, such as typing near normal speeds, while remaining immersed in VR.

10.4 APPLICATIONS & FUTURE WORK

Though this thesis has concerned itself with music retrieval and user engagement, many of the techniques developed are more broadly applicable. The use of generative models to infer music selections is used in both chapters 6 and 7, and given the simple interaction models used, result in emergent interaction dynamics that can be tuned to a particular user. Developing models of user input, conditioned upon user goals, allows for a user-centred approach to designing interactive systems. A belief about a user's goal can be inferred from their input, using a model trained to them. Such systems are well suited to tasks such as recommendation, which involve beliefs over collections rather than explicit selections. A key limitation however is that any such system will inherently be highly sensitive to a model that poorly fits actual user behaviour.

Another theme developed throughout this thesis is the use of information-theoretic measures to provide meta-analyses of features and services. This was used in section 5.2 to compare and rank music features from multiple sources. A related approach was applied in the evaluation of the BeoSound Moment in section 8.3, where the diversity of results from different recommender strategies was contrasted. This style of meta-analysis could provide a means of quantitatively comparing the offerings of recommender services, and how they enrich and grow a user's interaction with content. More generally, such measures could also be used to inform and encourage users to explore and seek out diverse experiences. This could apply not only in terms of music recommenders and the range of music listened to, but also for GPS-enabled devices encouraging users to explore new locations, or a calorie tracking mobile app rewarding a varied diet.

10.5 CONTROL

While this thesis has done much to motivate and explore designing for music engagement, the underlying principles being investigated in this work relate to the user's *control*. As the interactive systems available to users become increasingly intelligent and autonomous, making recommendations and inferring user intent, the issue of diminished control becomes increasingly pressing. Interfaces that afford various levels of engagement and control go some way towards addressing this issue, however, there remain further challenges. Where users do not understand or trust a recommender system, they will not be as empowered when handing over control to it. For example, a user may avoid listening to embarrassing tracks to avoid these tracks being recommended later in an inappropriate context. Similarly, recommenders may be biased, as in the case of tracks played by commercial radio being heavily influenced by record labels. Much further consideration is needed of the interaction and balance of control between user and agent. In particular, where users delegate control to an agent, the agent's (and its designer's) control and influence over the user must then be questioned.

MORAL HAZARD

Music retrieval systems are the gateway by which many users access their music. The issue of recommendation being biased, whether intentionally for commercial reasons or due to a biased classifier, raises a significant moral hazard. Users place trust in recommender systems, often using them to manage and access their music collections. Any bias in these systems not only has commercial implications, but also shapes the cultural exposure of users. In the discussion of music classification in chapter 2, it is observed that classifiers are confounded by equalisation and the process by which music is produced. Such a lack of generalisability does not only impact performance; where music classifiers are developed and trained on a corpus of music from large music labels, it is questionable whether their performance would generalise to independently produced music. As music listening becomes increasingly conditioned upon classifiers and recommender systems, this bias in music retrieval becomes an increasing concern. Music Information Retrieval has reached a stage where the work in field can have significant financial and cultural implications. It is thus important to be mindful of this moral hazard and to not consider performance measures alone.

10.6 SUMMARY

Users differ in how much they engage with their music, their retrieval systems and their interactions in general, investing corresponding amounts of effort and attention. This thesis has measured how users' behaviour varies with their engagement, and developed engagement–stratified user profiles. The design of interactive systems can be informed by this range of user engagement. This thesis has seen examples of retrieval systems affording a range of user engagement levels, and interactions that can adapt to the user's engagement context. This engagement-dependent approach to interaction design not only frees users from the issues of too-much-choice and disempowerment, but can be generalised, for example solving usability issues with Virtual Reality Head Mounted Displays. Engagement-dependent retrieval offers a mechanism for empowering users with both intelligent recommendation and control, with promising applications in future music retrieval systems.

A. PRESSURE SENSING

PRESSURE input can be used as a way for users to explicitly denote their engagement, using a metaphor of physical exertion as engagement. This is employed in chapter 7, where the user can set their level of engagement in a music retrieval interaction, with an adaptive interface responding accordingly. The addition of a pressure input modality to an interaction requires the use of a pressure sensor and the details of the implementation used in this thesis are given in this chapter. Pressure input is noisy and, if mapped directly to the interaction, will be difficult for users to control. A Fitts' law (1954) study is performed in order to characterise pressure input and users' ability to control it, and in particular to investigate the utility of low-pass filtering in improving the usability of pressure input.

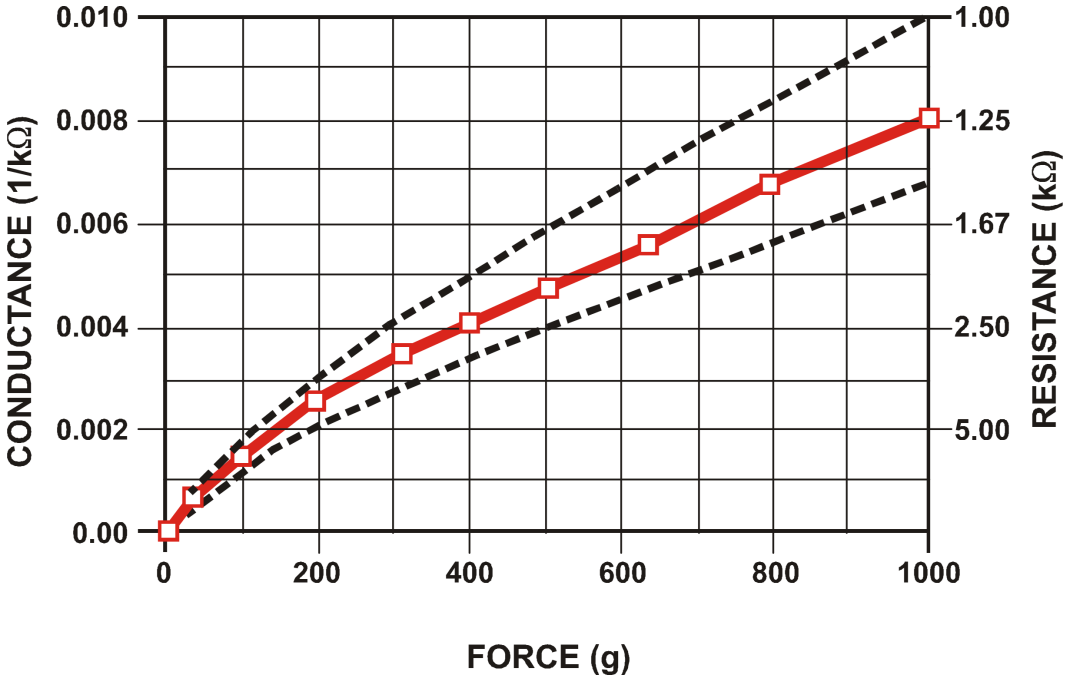

Figure A.1: Characteristics of the Interlink Force Sensitive Resistor (FSR) used for pressure-sensing. At low levels of force (within human ranges) the relationship with conductance is approximately linear, and thus usable for Human-Computer Interaction. Figure taken from the Interlink FSR integration guide.[1]

A.1 HARDWARE

The force applied by the user was measured using an Interlink Electronics Force Sensitive Resistor (FSR)[1]. The conductance of this sensor is linearly proportional to force within human ranges, as shown in Figure A.1. As current flow is linear with conductance (and thus the applied force), a transimpedance amplifier or 'current-to-voltage converter' is used to obtain a voltage that represents force (depicted in Figure A.2). This approach takes advantage of the linear force–conductance property of the sensor at human input levels, avoiding the issues of a non-linear input modality. This setup resulted in a pressure space of 10N, which was used in developing the prototype system in chapter 7. In this implementation, an mbed[2] microcontroller was used to connect the FSR to a Microsoft Surface Pro tablet running Windows 8, which ran the prototype software. The prototype software made use of a low-pass filter to make the pressure input more usable, and the selection of the filter properties is explored in this chapter. This exploration uses the same hardware setup, with an experimental interface configured for a Fitts (1954) targeting study.

[1] `https://www.sparkfun.com/datasheets/Sensors/Pressure/fsrguide.pdf`
[2] `http://www.mbed.org` (06/12/14)

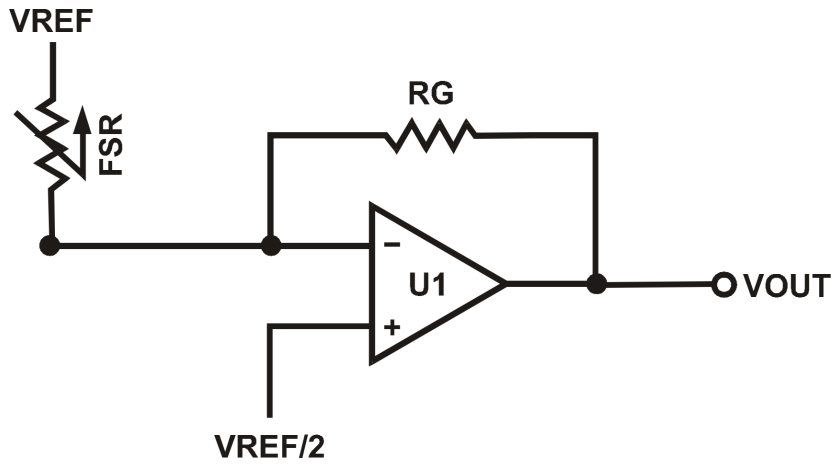

Figure A.2: The transimpedance amplifier or 'current to voltage' converter produces a voltage output that is linear with the current input. The conductance (and thus current) of the FSR is approximately linear with force across the range of human input, yielding a linear force–voltage relationship. Figure taken from the Interlink FSR integration guide.[1]

A.2 LOW PASS FILTER

Initial prototyping with the pressure sensor used the raw measured values. These showed a high degree of noise, rendering control difficult. While the sensor itself is subject to some thermal noise and hysteresis, the majority of the noise observed is likely due to muscle tremor. In order to mitigate this issue, the use of a low-pass filter is explored. A low-pass filter attenuates the high frequency content of a signal – in this case, the erratic movement due to muscle tremor. The low frequency content remains, i.e. the user's controlled movement through the pressure input space. The selection of appropriate filter parameters is essential, as filtering too aggressively would make the on-screen feedback of input pressure noticeably delayed from the user's input. In extreme cases, rapid changes in input pressure with a slow on-screen response could make the interaction difficult for the user to control.

The implementation of the low-pass filter uses the simple recurrence relation form, also known as the exponentially weighted moving average:

$$y_i = y_{i-1} + \alpha(x_i - y_{i-1}),$$

where α is a smoothing factor within the range: $0 < \alpha < 1$. It can be related to an RC low-pass filter time constant τ, where f_s is the sampling frequency. In this work, $f_s = 325Hz$.

$$\alpha = \frac{1}{f_s\,\tau + 1}.$$

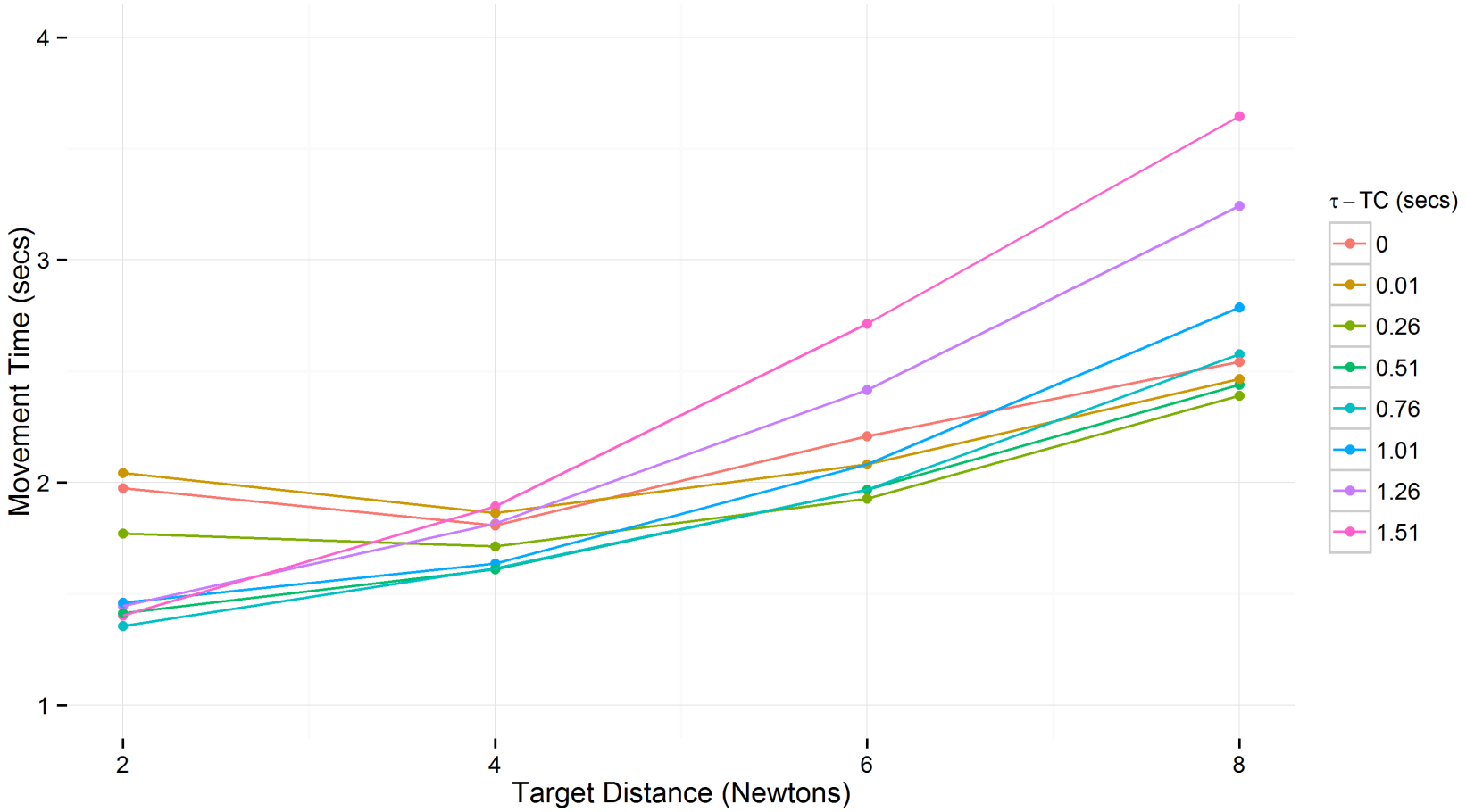

Figure A.3: Results from a Fitts' law study of pressure interaction across a range of low-pass filter *time constant* τ values. The reciprocal of the slope of the line indicates the interaction throughput, thus a less steep line is preferable. Note that for the lowest τ values at the smallest distance, users struggled to control the pressure and Fitts' law does not apply.

A.3 PRESSURE INPUT DYNAMICS STUDY

A Fitts (1954) targeting study was conducted to identify an appropriate filter time constant τ for Human Computer Interaction. Such a study involves asking participants to repeatedly select targets at a range of distances, in this case with the use of the pressure input. The *throughput* (in bits per second) of an interface can be determined by capturing the *movement time* (MT) of the user's targeting, as well as adjusting for their accuracy (Zhai, 2004). Given the standard Fitts' law form:

$$\text{MT} = a + b\,\text{ID},$$

MT is linear with the *index of difficulty* (ID), which is a function of target distance and width. Throughput (TP) is given in ISO 9241-9 using the slope-inverse method:

$$\text{TP} = \frac{\text{ID}_e}{\text{MT}},$$

making the assumption of a zero intercept, i.e. $a = 0$. ID_e reflects the *empirical index of difficulty*, with target width having been adjusted to fit the error distribution observed. The zero intercept assumption is unnecessary however, by applying linear regression with $\text{TP} = \frac{1}{b}$ (see Zhai (2004) for a discussion on throughput calculation).

SELECTING A FILTER TIME CONSTANT

16 users were recruited to perform the targeting study. The study comprised of 4 target distances $\times$ 15 repetitions $\times$ 8 time constant τ values $\times$ 16 participants, giving 7680 trials (an additional 480 were used as training trials). The average movement time per target distance for each filter is depicted in Figure A.3. For the filters $\tau \leq 0.26$, the MT at the $2N$ distance was higher than would be predicted by the linear Fitts' law relationship. Users struggled to control the pressure level at the small distance with these least amounts of filtering.

The results of the Fitts' law regression are given in table A.4. The r^2 values for the first three filters shows they do not conform to Fitts' law and thus their TP is invalid. The next filter, $\tau = 0.51$, thus has the highest usable TP and is adopted for the work in this thesis.

τ	0.00	0.01	0.26	0.51	0.76	1.01	1.26	1.51
a	2.34	2.28	2.18	2.11	2.04	2.10	2.26	2.43
TP	4.70	5.81	4.17	2.65	1.88	1.97	1.44	1.04
r^2	0.39	0.36	0.58	0.92	0.84	0.88	0.94	0.92

Table A.4: Pressure input throughput from Fitts' law regression

B. Combined Music Engagement Questionnaire

THE following music engagement and use questionnaire is based on similar questionnaires in Greasley and Lamont (2011); Krause et al. (2014). These previous questionnaires were combined, and the wording of some questions updated to reflect digital music listening. This questionnaire begins with two questions from Krause et al.'s questionnaire, asking participants to tick any selection methods and devices that they use when listening to music. The remaining questions are adapted from Greasley and Lamont's questionnaire and prompt participants to respond using a five point Likert scale.[1]

[1] Original form as given to users: `https://goo.gl/TTfXeM`

QUESTIONNAIRE

The aim of this questionnaire is to gather data on people's music listening behaviour. You will be asked to provide your Last.fm username so that your publicly shared Scrobbles can be analysed. Completing this questionnaire is taken as consent for this analysis.

Please enter your Last.fm username:

Username

We will only analyse your public Scrobbles to identify your music-listening behaviour.

Think about the last few times you listened to music and tick the selection methods you used.

- ☐ I did not have any control
- ☐ It was performed live at the time
- ☐ Watched TV
- ☐ Personal premade playlist
- ☐ Specific artist
- ☐ Premade playlist - by someone else
- ☐ Specific song
- ☐ Listened to the radio
- ☐ Specific album
- ☐ Someone I was with chose
- ☐ Random/Shuffle
- ☐ Computer Recommendation
- ☐ Website streaming
- ☐ Other:

  Please state

Please tick the devices you often use to listen to music.

- ☐ Mobile MP3
- ☐ Mobile Phone
- ☐ Mobile CD
- ☐ Computer - own music
- ☐ Computer - streamed music
- ☐ Computer - music store in the cloud
- ☐ Stereo - MP3
- ☐ Stereo - CD
- ☐ Radio
- ☐ TV
- ☐ In public - live
- ☐ In public - recorded

How much control do you like to have over the music you listen to?

☐ I generally listen to whatever is played

☐ I listen to a particular radio channel or recommendations in a style of music I like

☐ Sometimes I like to choose, sometimes I'll listen to radio or recommendations

☐ I generally prefer listening to music I've chosen e.g. in a playlist I made

☐ I like to have full control over which album or song I'm listening to

How much music do you have in your collection (in iTunes, MP3 player, Spotify etc)?

☐ Up to 5 albums or about 50 MP3s

☐ Up to 25 albums or about 250 MP3s

☐ Up to 125 albums or about 1250 MP3s

☐ Up to 5000 albums or about 50,000 MP3s

☐ More than 5000 albums or 50,000 MP3s

What best describes your music collection?

☐ All my collection is organised by hand in a certain way

☐ My collection is broadly categorised by hand (e.g. genre or mood playlists/folders)

☐ My music system organises my collection for me

☐ There is no specific organisation to my music collection

☐ I can never find anything when I want it!

Can you remember the first album you bought?

☐ Yes, I can remember what it was and exactly when and where I bought it

☐ I can remember what it was, but not exactly when I bought it

☐ Not off hand, I probably could if I thought about it

☐ I doubt it

☐ I have no idea

Why did you make your last music purchase?

☐ I had to have it, I heard it and I just had to go and buy it

☐ I knew I would like it, because I've built up a knowledge of what I like

☐ I'd heard a couple of the tracks I liked so bought it to see if the whole thing was good

☐ It was recommended to me, so I thought I'd give it a go

☐ I can't remember

Lyrics in music, which best describes you?

☐ I never really listen to the lyrics in songs

☐ With some of my favourite music I know the lyrics, but otherwise I don't really pay attention to the lyrics in songs

☐ I like to know the lyrics so I can sing along

☐ I have to know the lyrics because it's central to understanding what the artist is trying to convey

☐ I have to know the lyrics because I don't want to be singing along to something that might be at odds with my beliefs

C. BUILDING THIS THESIS

This thesis is made available as an open source, reproducible document. The data and code used to generate the analyses and graphs within this document are made available for reuse and inspection under the MIT open source license. The code is structured as an example of *literate programming*, interleaving the LaTeX and R code used in making the arguments of this thesis. Instructions for building this thesis are included in this chapter, as well as a description of the included data.

The commands in listing C.1 will download and build this document from source, as well as performing the analyses and meeting dependencies, such as downloading the SPUD dataset (chapter 4) and building required R packages. It is assumed that R and LaTeX environments are installed, as well as the *make*, *wget* and *git* utilities.

Listing C.1: Building the thesis

```
# git clone http://www.github.com/DCBoland/thesis.git
# cd thesis
# make
```

The resulting document can be relatively large. If *ghostscript* is installed, the command in listing C.2 is included as a convenience function to optimise the document size.

Listing C.2: Reducing the document size

```
# make compress
```

The information-theoretic feature selection in section 5.2 made use of proprietary music features. This music feature data is not distributed with the thesis. The corresponding figure (Figure 5.2) is instead distributed as a static image. The code used for the analysis is included, however, with a cache file used to prevent its execution during compilation.

C.1 R ENVIRONMENT & PACKAGES

This thesis has made use of R and a number of R packages, which are the result of work by a great many people. The code and included comments indicate which packages were used for a given analysis. Many package authors provide details for citations, and their efforts are acknowledged here:

R (R Core Team, 2014), knitr (Xie, 2014), dplyr (Wickham and Francois, 2015), tidyr (Wickham, 2014), caret (Kuhn, 2015), RSQLite (Wickham et al., 2014), changepoint (Rebecca Killick and Haynes, 2014), ggplot2 (Wickham, 2009), infotheo (Meyer, 2014), likert (Bryer and Speerschneider, 2014), mvtnorm (Genz and Bretz, 2009), polycor (Fox, 2010), nlme (Pinheiro et al., 2015), multcomp (Hothorn et al., 2008), survival (Therneau and Grambsch, 2000).

C.2 DATABASE SCHEMA

The SPUD dataset is distributed as a SQLite file. The data is normalised in fifth normal form, with mutually associated TRACK, ALBUM and ARTIST tables, as well as corresponding Spotify IDs for each record. The TRACK table also contains track metadata acquired from Spotify, such as popularity and duration. The Last.fm playlist data is included, as the LASTFMPLAYLISTS and LASTFMUSERS tables. A cross-reference table, LASTFMPLAYLISTSTRACKS, is used to associate each playlist with the contained tracks. Some users' listening histories are also provided, using the cross-reference table LASTFMTRACKLISTENS.

For all of the other studies in this thesis, the required data is included in CSV format in the *data* directory.

BIBLIOGRAPHY

AUCOUTURIER, J. J. AND PACHET, F. Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio*, 1(1): 1–13 (2004).

AZZOPARDI, L. The economics in interactive information retrieval. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, SIGIR, pages 15–24. ACM, Beijing, China (2011).

AZZOPARDI, L. Modelling interaction with economic models of search. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, SIGIR, pages 3–12. ACM, Gold Coast, Australia (2014).

BARRETT, J. AND KRUEGER, H. Performance effects of reduced proprioceptive feedback on touch typists and casual users in a typing task. *Behaviour & Information Technology*, 13(6): 373–381 (1994).

BENFORD, S., GREENHALGH, C., REYNARD, G., BROWN, C., AND KOLEVA, B. Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(3): 185–223 (1998).

BERGSTRA, J., CASAGRANDE, N., ERHAN, D., ECK, D., AND KÉGL, B. Aggregate features and AdaBoost for music classification. *Machine learning*, 65(2-3): 473–484 (2006).

BERTIN-MAHIEUX, T., ELLIS, D. P., WHITMAN, B., AND LAMERE, P. The Million Song Dataset. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pages 591–596. Miami, Florida, USA (2011).

BILLINGHURST, M. AND KATO, H. Collaborative mixed reality. In *Proceedings of International Symposium on Mixed Reality*, ISMR, pages 261–284. ACM, Yokohama, Japan (1999).

BÖCK, S., KREBS, F., AND SCHEDL, M. Evaluating the online capabilities of onset detection methods. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 49–54. Porto, Portugal (2012).

BOLAND, D. AND MURRAY-SMITH, R. Finding *my* beat: personalised rhythmic filtering for mobile music interaction. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI, pages 21–30. ACM, Munich, Germany (2013).

BOLAND, D. AND MURRAY-SMITH, R. Information-theoretic measures of music listening behaviour. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 561–566. Taipei, Taiwan (2014).

BORN, G. AND HESMONDHALGH, D. *Western music and its others: Difference, representation, and appropriation in music*. University of California Press (2000).

BOZKURT, B., AYANGIL, R., AND HOLZAPFEL, A. Computational analysis of turkish makam music: Review of state-of-the-art and challenges. *Journal of New Music Research*, 43(1): 3–23 (2014).

BROWN, G., POCOCK, A., ZHAO, M.-J., AND LUJÁN, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13: 27–66 (2012).

BRUDER, G., STEINICKE, F., VALKOV, D., AND HINRICHS, K. Augmented virtual studio for architectural exploration. In *Proceedings of Virtual Reality International Conference*, VRIC, pages 1–10. Laval, France (2010).

BRYER, J. AND SPEERSCHNEIDER, K. *likert: Functions to analyze and visualize likert type items* (2014). R package version 1.2.

CARVALHO, F. G., TREVISAN, D. G., AND RAPOSO, A. Toward the design of transitional interfaces: an exploratory study on a semi-immersive hybrid user interface. *Virtual Reality*, 16(4): 271–288 (2012).

CASEY, M., VELTKAMP, R., AND GOTO, M. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4): 668–696 (2008).

CELMA, Ò. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer (2010).

CELMA, Ò. AND CANO, P. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, KDD. ACM, Las Vegas, USA (2008).

CLAWSON, J., LYONS, K., STARNER, T., AND CLARKSON, E. The impacts of limited visual feedback on mobile text entry for the twiddler and mini-qwerty keyboards. In *Proceedings of the International Symposium on Wearable Computers*, ISWC, pages 170–177. IEEE, Washington DC, USA (2005).

COCKBURN, A. AND FIRTH, A. Improving the acquisition of small targets. In *People and Computers XVII: Designing for Society*, pages 181–196. Springer (2004).

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2): 187–220 (1972).

CRAFT, A., WIGGINS, G., AND CRAWFORD, T. How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 74–76. Vienna, Austria (2007).

CROSSAN, A. AND MURRAY-SMITH, R. Rhythmic interaction for song filtering on a mobile device. In *Proceedings of The International Conference on Haptic and Audio Interaction Design*, HAID, pages 45–55. Springer, Glasgow, UK (2006).

CSIKSZENTMIHALYI, M., ABUHAMDEH, S., AND NAKAMURA, J. Flow. In *Flow and the Foundations of Positive Psychology*, pages 227–238. Springer (2014).

CUNNINGHAM, S. J., JONES, S., AND JONES, M. Organizing digital music for use: an examination of personal music collections. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR. Barcelona, Spain (2004).

CUNNINGHAM, S. J., REEVES, N., AND BRITLAND, M. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the Joint Conference on Digital libraries*, JCDL. IEEE, Texas, USA (2003).

DAVIS, L., ROLLAND, J., HAMZA-LUP, F., HA, Y., NORFLEET, J., AND IMIELINSKA, C. Z. Enabling a continuum of virtual environment experiences. *IEEE Computer Graphics and Applications*, 23(2): 10–12 (2003).

DOWLING, W. J. Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4): 341–354 (1978).

DOWNIE, J. S. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37(1): 295–340 (2003).

DOWNIE, J. S. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12): 795–825 (2006).

DRAKE, C. AND BERTRAND, D. The Quest for Universals in Temporal Processing in Music. *Annals of the New York Academy of Science*, 930: 17–27 (2001).

DRAKE, C. AND EL HENI, J. B. Synchronizing with music: Intercultural differences. *Annals of the New York Academy of Sciences*, 999: 429–437 (2003).

DUPRET, G. AND LALMAS, M. Absence time and user engagement: Evaluating ranking functions. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM, pages 173–182. ACM, Rome, Italy (2013).

EGERMANN, H., FERNANDO, N., CHUEN, L., AND MCADAMS, S. Music induces universal emotion-related psychophysiological responses: Comparing canadian listeners to congolese pygmies. *Frontiers in Psychology*, 5(1341): 173–182 (2015).

FABBRI, F. What kind of music? *Popular Music*, 2: 131–143 (1982).

FITTS, P. M. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6): 381 (1954).

FLEMISCH, O., ADAMS, A., CONWAY, S. R., GOODRICH, K. H., PALMER, M. T., AND SCHUTTE, P. C. NASA/TM–2003–212672 The H-Metaphor as a Guideline for Vehicle Automation and Interaction. Technical report, NASA (2003).

FLEXER, A. On inter-rater agreement in audio music similarity. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 245–250. Taipei, Taiwan (2014).

FOX, J. *polycor: Polychoric and Polyserial Correlations* (2010). R package version 0.7-8.

FUTRELLE, J. AND DOWNIE, J. S. Interdisciplinary Research Issues in Music Information Retrieval: ISMIR 2000 - 2002. *Journal of New Music Research*, 32(2): 121–131 (2003).

GENZ, A. AND BRETZ, F. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer, Heidelberg (2009).

GHIAS, A., LOGAN, J., CHAMBERLIN, D., AND SMITH, B. C. Query by humming: musical information retrieval in an audio database. In *Proceedings of the International Conference on Multimedia*, MM, pages 231–236. ACM, San Francisco, California, USA (1995).

GHOMI, E., FAURE, G., HUOT, S., AND CHAPUIS, O. Using rhythmic patterns as an input method. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 1253–1262. ACM, Austin, Texas, USA (2012).

GJERDINGEN, R. O. AND PERROTT, D. Scanning the Dial: The Rapid Recognition of Music Genres. *Journal of New Music Research*, 37(2): 93–100 (2008).

GONZÁLEZ, G., MOLINA, J. P., GARCÍA, A. S., MARTÍNEZ, D., AND GONZÁLEZ, P. Evaluation of text input techniques in immersive virtual environments. In MACÍAS, J. A. ET AL. (editors), *New Trends on Human–Computer Interaction*, pages 109–118. Springer, London (2009).

GRAHAM, R. L., KNUTH, D. E., AND PATASHNIK, O. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Boston, MA, USA, 2nd edition (1994).

GRASSET, R., LAMB, P., AND BILLINGHURST, M. Evaluation of mixed-space collaboration. In *Proceedings of the International Symposium on Mixed and Augmented Reality*, ISMAR, pages 90–99. IEEE/ACM (2005).

GREASLEY, A. E. *Engagement with music in everyday life: and in-depth study of adults' musical preferences and listening behaviours*. Ph.D. thesis, University of Keele (2008).

GREASLEY, A. E. AND LAMONT, A. Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae*, 15(1): 45–71 (2011).

HANNA, P. AND ROBINE, M. Query by tapping system based on alignment algorithm. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, ICASSP, pages 1881–1884. IEEE, Taipei, Taiwan (2009).

HELMHOLTZ, H. L. *On the sensations of tone as a physiological basis for the theory of music*. Longmans, Green, London, 3rd edition (1895).

HICK, W. E. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1): 11–26 (1952).

HIGGINS, K. M. *The Music Between Us: Is Music a Universal Language?* The University of Chicago Press, Chicago (2012).

HOTHORN, T., BRETZ, F., AND WESTFALL, P. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3): 346–363 (2008).

HU, X. AND LIU, J. Evaluation of music information retrieval : Towards a user-centered approach. In *Proceedings of the Workshop on Human-Computer Interaction and Information Retrieval*, HCIR. New Brunswick, Canada (2010).

HYMAN, R. Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45(3): 188–196 (1953).

IYENGAR, S. S. AND LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6): 995–1006 (2000).

JANG, J., LEE, H., AND YEH, C.-H. Query by tapping: A new paradigm for content-based music retrieval from acoustic input. In *Advances in Multimedia Information Processing–PCM*, volume 2195 of *LNCS*, pages 590–597. Springer (2001).

JIANG, J., AWADALLAH, A. H., SHI, X., AND WHITE, R. W. Understanding and predicting graded search satisfaction. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM, pages 57–66. ACM, Shanghai, China (2015).

KABER, D. B. AND ZHANG, T. Human factors in virtual reality system design for mobility and haptic task performance. *Reviews of Human Factors and Ergonomics*, 7: 323–366 (2011).

KANESHIRO, B., KIM, H.-S., HERRERA, J., OH, J., BERGER, J., AND SLANEY, M. Qbt-extended: An annotated dataset of melodically contoured tapped queries. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 329–334. Curitiba, Brazil (2013).

KAPLAN, E. L. AND MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): 457–481 (1958).

KAPOOR, K., SUN, M., SRIVASTAVA, J., AND YE, T. A hazard based approach to user return time prediction. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, KDD, pages 1719–1728. ACM, New York, USA (2014).

KARAGANIS, J. AND RENKEMA, L. *Copy Culture in the US and Germany*. American Assembly (2013).

KASSLER, M. Toward musical information retrieval. *Perspectives of New Music*, 4(2): 59–67 (1966).

KILLICK, R., FEARNHEAD, P., AND ECKLEY, I. A. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500): 1590–1598 (2012).

KLYUBIN, A. S., POLANI, D., AND NEHANIV, C. L. Empowerment: A universal agent-centric measure of control. In *The Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE, Edinburgh, UK (2005).

KRAUSE, A., NORTH, A., AND HEWITT, L. Music selection behaviors in everyday listening. *Journal of Broadcasting & Electronic Media*, 58(2): 306–323 (2014).

KRISTENSSON, P. O. AND VERTANEN, K. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings of the International Conference on Intelligent User Interfaces*, IUI, pages 29–32. ACM, Lisbon, Portugal (2012).

KUHN, M. *caret: Classification and Regression Training* (2015). R package version 6.0-41.

KUHN, M., WATTENHOFER, R., WIRZ, M., FLUCKIGER, M., AND TROSTER, G. Sensing dance engagement for collaborative music control. In *Proceedings of the International Semantic Web Conference*, ISWC, pages 51–54. IEEE, Bonn, Germany (2011).

LAFFERTY, J. AND ZHAI, C. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, SIGIR, pages 111–119. ACM, New Orleans, Louisiana, USA (2001).

LALMAS, M., O'BRIEN, H., AND YOM-TOV, E. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4): 1–132 (2014).

LANTZ, V. AND MURRAY-SMITH, R. Rhythmic interaction with a mobile device. In *Proceedings of the Nordic Conference on Human-Computer Interaction*, NordiCHI, pages 97–100. ACM, Tampere, Finland (2004).

LAPLANTE, A. *Everyday Life Music Information-Seeking Behaviour of Young Adults: An Exploratory Study*. Ph.D. thesis, McGill University, Quebec (2008).

LEE, J., DOWNIE, J., AND CUNNINGHAM, S. Challenges in cross-cultural/multilingual music information seeking. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 1–7. London, UK (2005).

LEE, J. H. AND CUNNINGHAM, S. J. The impact (or non-impact) of user studies in music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 391–396. Porto, Portugal (2012).

LEE, J. H. AND DOWNIE, J. S. Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 441–446. Barcelona, Spain (2004).

LEHMANN, J., LALMAS, M., YOM-TOV, E., AND DUPRET, G. Models of user engagement. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, UMAP, pages 164–175. Springer, Berlin, Heidelberg (2012).

LEMAN, M., STYNS, F., AND BERNARDINI, N. Sound, sense and music mediation: a historical-philosophical perspective. In *Sound to sense, sense to sound: a state of the art in sound and music computing*, pages 15–46. Logos (2008).

LEONG, T., VETERE, F., AND HOWARD, S. The serendipity shuffle. In *Proceedings of the Australian Computer-Human Interaction Conference*, OzCHI, pages 25–28. ACM, Canberra, Australia (2005).

LONG, J. Visual feedback and skilled keying: Differential effects of masking the printed copy and the keyboard. *Ergonomics*, 19(1): 93–110 (1976).

LONSDALE, A. J. AND NORTH, A. C. Why do we listen to music? a uses and gratifications analysis. *British Journal of Psychology*, 102: 108–134 (2011).

MANABE, H. AND FUKUMOTO, M. Headphone taps: a simple technique to add input function to regular headphones. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI, pages 177–179. ACM, San Francisco, California, USA (2012).

MASRI, P. *Computer modelling of sound for transformation and synthesis of musical signals*. Ph.D. thesis, University of Bristol (1996).

MCFEE, B. AND LANCKRIET, G. Hypergraph models of playlist dialects. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 343–348. Porto, Portugal (2012).

MCGILL, M., BOLAND, D., MURRAY-SMITH, R., AND BREWSTER, S. A. A dose of reality: Overcoming usability challenges in VR head-mounted displays. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 2143–2152. ACM, Seoul, Korea (2015).

MEYER, P. E. *infotheo: Information-Theoretic Measures* (2014). R package version 1.2.0.

MILGRAM, P. AND COLQUHOUN, H. A taxonomy of real and virtual world display integration. In OHTA, Y. AND TAMURA, H. (editors), *Mixed reality: Merging real and virtual worlds*, pages 5–30. Springer, Berlin (1999).

MILGRAM, P. AND KISHINO, F. A Taxonomy Of Mixed Reality Visual Displays. *Institute of Electronics, Information and Communication Engineers Transactions on Information and Systems*, 77(12): 1321–1329 (1994).

MONAHAN, C. B. AND CARTERETTE, E. C. Pitch and duration as determinants of musical space. *Music Perception*, 3(1): 1–32 (1985).

MONGEAU, M. AND SANKOFF, D. Comparison of musical sequences. *Computers and the Humanities*, 24(3): 161–175 (1990).

NORMAN, D. A. *The psychology of everyday things*. Basic books (1988).

OBENDORF, H. *Minimalism*. Springer (2009).

O'BRIEN, H. L. AND TOMS, E. G. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6): 938–955 (2008).

O'BRIEN, H. L. AND TOMS, E. G. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1): 50–69 (2010).

OULASVIRTA, A., TAMMINEN, S., ROTO, V., AND KUORELAHTI, J. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 919–928. ACM, Oregon, USA (2005).

PANAGAKIS, Y. AND KOTROPOULOS, C. Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, ICASSP, pages 249–252. IEEE, Dallas, Texas, USA (2010).

PERLMAN, M. AND KRUMHANSL, C. L. An Experimental Study of Internal Interval Standards in Javanese and Western Musicians. *Music Perception*, 14(2): 95–116 (1996).

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D., AND R CORE TEAM. *nlme: Linear and Nonlinear Mixed Effects Models* (2015). R package version 3.1-120.

POHL, H. AND MURRAY-SMITH, R. Focused and casual interactions: Allowing users to vary their level of engagement. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 2223–2232. ACM, San Jose, USA (2013).

QUIÑONES, M. Listening in Shuffle Mode. *Lied und populäre Kultur/Song and Popular Culture*, 52: 11–22 (2007).

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2014).

REBECCA KILLICK, I. E. AND HAYNES, K. *changepoint: An R package for changepoint analysis* (2014). R package version 1.1.5.

REKIMOTO, J. AND NAGAO, K. The world through the computer: Computer augmented interaction with real world environments. In *Proceedings of the User Interface Software and Technology Symposium*, UIST, pages 29–36. ACM, Pittsburgh, Pennsylvania, USA (1995).

REPETTO, R. C. AND SERRA, X. Creating a corpus of Jingju (Beijing opera) music and possibilities for melodic analysis. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 313–318. Taipei, Taiwan (2014).

RICO, J. AND BREWSTER, S. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 887–896. ACM, Atlanta, Georgia, USA (2010).

ROBERTSON, S. On the history of evaluation in IR. *Journal of Information Science*, 34(4): 439–456 (2008).

ROSINSKI, R. R., CHIESI, H., AND DEBONS, A. Effects of amount of visual feedback on typing performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 24(1): 195–199 (1980).

SAKAI, T. AND DOU, Z. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, SIGIR, pages 473–482. ACM, Dublin, Ireland (2013).

SAPONAS, T. S., HARRISON, C., AND BENKO, H. Pockettouch: through-fabric capacitive touch input. In *Proceedings of the User Interface Software and Technology Symposium*, UIST, pages 303–308. ACM, Santa Barbara, California, USA (2011).

SCHEDL, M. AND FLEXER, A. Putting the user in the center of music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 385–390. Porto, Portugal (2012).

SCHEDL, M., FLEXER, A., AND URBANO, J. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3): 523–539 (2013).

SCHEDL, M., STOBER, S., GÓMEZ, E., ORIO, N., AND LIEM, C. C. S. User-Aware Music Retrieval. In MÜLLER, M., GOTO, M., AND SCHEDL, M. (editors), *Multimodal Music Processing*, pages 135–156. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik (2012).

SCHEIBEHENNE, B., GREIFENEDER, R., AND TODD, P. M. What moderates the too-much-choice effect? *Journal of Psychology & Marketing*, 26(3): 229–253 (2009).

SCHUBERT, T., FRIEDMANN, F., AND REGENBRECHT, H. The experience of presence: Factor analytic insights. *Presence: Teleoperators and virtual environments*, 10(3): 266–281 (2001).

SCHWARTZ, B. *The Paradox of Choice: Why More Is Less*. Harper Perennial (2005).

SCHWARTZ, B., WARD, A., MONTEROSSO, J., LYUBOMIRSKY, S., WHITE, K., AND LEHMAN, D. R. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5): 1178–1197 (2002).

SEPPÄNEN, J. Tatum grid analysis of musical signals. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 131–134. IEEE (2001).

SERRA, X. A multicultural approach in music information research. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 151–156. Miami, Florida, USA (2011).

SEYERLEHNER, K., WIDMER, G., AND KNEES, P. A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *Proceedings of the International Workshop on Adaptive Multimedia Retreival*, AMR, pages 118–131. Springer, Linz, Austria (2010).

SHAO, B. *User-centric Music Information Retrieval*. Ph.D. thesis, Florida International University (2011).

SHNEIDERMAN, B. AND PLAISANT, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Addison Wesley, 4 edition (2004).

SIMON, H. A. *Reason in human affairs*. Stanford University Press (1990).

SLATER, M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535): 3549–3557 (2009).

SMITH, T. F. AND WATERMAN, M. S. Identification of common molecular subsequences. *Molecular Biology*, 147(1): 195–197 (1981).

SOUKOREFF, R. W. AND MACKENZIE, I. S. Metrics for text entry research. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI, pages 113–120. ACM, Florida, USA (2003).

STEINICKE, F., BRUDER, G., ROTHAUS, K., AND HINRICHS, K. A virtual body for augmented virtuality by chroma-keying of egocentric videos. In *Symposium on 3D User Interfaces*, 3DUI, pages 125–126. IEEE, Lafayette, LA, USA (2009).

STOBER, S. *Adaptive Methods for User-Centered Organization of Music Collections*. Ph.D. thesis, Otto-von-Guericke-Universitï£¡t Magdeburg (2011).

STOWELL, D. AND PLUMBLEY, M. Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the International Computer Music Conference*, ICMC. Copenhagen, Denmark (2007).

STURM, B. L. An analysis of the GTZAN music genre dataset. In *Proceedings of the International Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM, pages 7–12. ACM, Nara, Japan (2012).

STURM, B. L. A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, 16(6): 1636–1644 (2014a).

STURM, B. L. A survey of evaluation in music genre recognition. In *Proceedings of the International Workshop on Adaptive Multimedia Retreival*, AMR, pages 29–66. Springer, Copenhagen, Denmark (2014b).

TECCHIA, F., AVVEDUTO, G., BRONDI, R., CARROZZINO, M., BERGAMASCO, M., AND ALEM, L. I'm in VR!: Using your own hands in a fully immersive MR system. In *Proceedings of the Symposium on Virtual Reality Software and Technology*, VRST, pages 73–76. ACM, Edinburgh, UK (2014).

THERNEAU, T. M. AND GRAMBSCH, P. M. *Modeling Survival Data: Extending the Cox Model*. Springer, New York (2000).

TREHUB, S. E. Human processing predispositions and musical universals. In WALLIN, N. L., MERKER, B., AND BROWN, S. (editors), *The Origins of Music*, chapter 23, pages 427–448. MIT Press (2000).

TZANETAKIS, G. AND COOK, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5): 293–302 (2002).

VAN BAREN, J. AND IJSSELSTEIJN, W. Measuring presence: A guide to current measurement approaches. Technical report, OmniPres project (2004).

VENNA, J., PELTONEN, J., NYBO, K., AIDOS, H., AND KASKI, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11: 451–490 (2010).

VINH, N. X., EPPS, J., AND BAILEY, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11: 2837–2854 (2010).

VOORHEES, E. M. The trec-8 question answering track report. In *Proceedings of the Text Retrieval Conference*, TREC, pages 77–82. Gaithersburg, Maryland, USA (1999).

WANG, A. ET AL. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 7–13. Baltimore, Maryland, USA (2003).

WEIGL, D. M. AND GUASTAVINO, C. User studies in the music information retrieval literature. In *Proceedings of the International Conference on Music Information Retrieval*, ISMIR, pages 335–340. Miami, Florida, USA (2011).

WICKHAM, H. *ggplot2: elegant graphics for data analysis*. Springer New York (2009).

WICKHAM, H. *tidyr: Easily Tidy Data with spread() and gather() Functions.* (2014). R package version 0.2.0.

WICKHAM, H. AND FRANCOIS, R. *dplyr: A Grammar of Data Manipulation* (2015). R package version 0.4.1.

WICKHAM, H., JAMES, D. A., AND FALCON, S. *RSQLite: SQLite Interface for R* (2014). R package version 1.0.0.

WIGGINS, G. Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music. In *Proceedings of the International Symposium on Multimedia*, ISM, pages 477–482. San Diego, California, USA (2009).

WILSON, M. L. AND ELSWEILER, D. Casual-leisure searching: the exploratory search scenarios that break our current models. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR (2010).

WOBBROCK, J. O. Tapsongs: tapping rhythm-based passwords on a single binary sensor. In *Proceedings of the User Interface Software and Technology Symposium*, UIST, pages 93–96. ACM, Victoria, BC, Canada (2009).

WOLFF, D. AND WEYDE, T. Learning music similarity from relative user ratings. *Information retrieval*, 17(2): 109–136 (2014).

XIE, Y. knitr: A comprehensive tool for reproducible research in R. In STODDEN, V., LEISCH, F., AND PENG, R. D. (editors), *Implementing Reproducible Computational Research*. Chapman and Hall/CRC (2014). ISBN 978-1466561595.

ZHAI, S. Characterizing computer input with Fitts' Law parameters: The information and non-information aspects of pointing. *International Journal of Human–Computer Studies*, 61(6): 791–809 (2004).